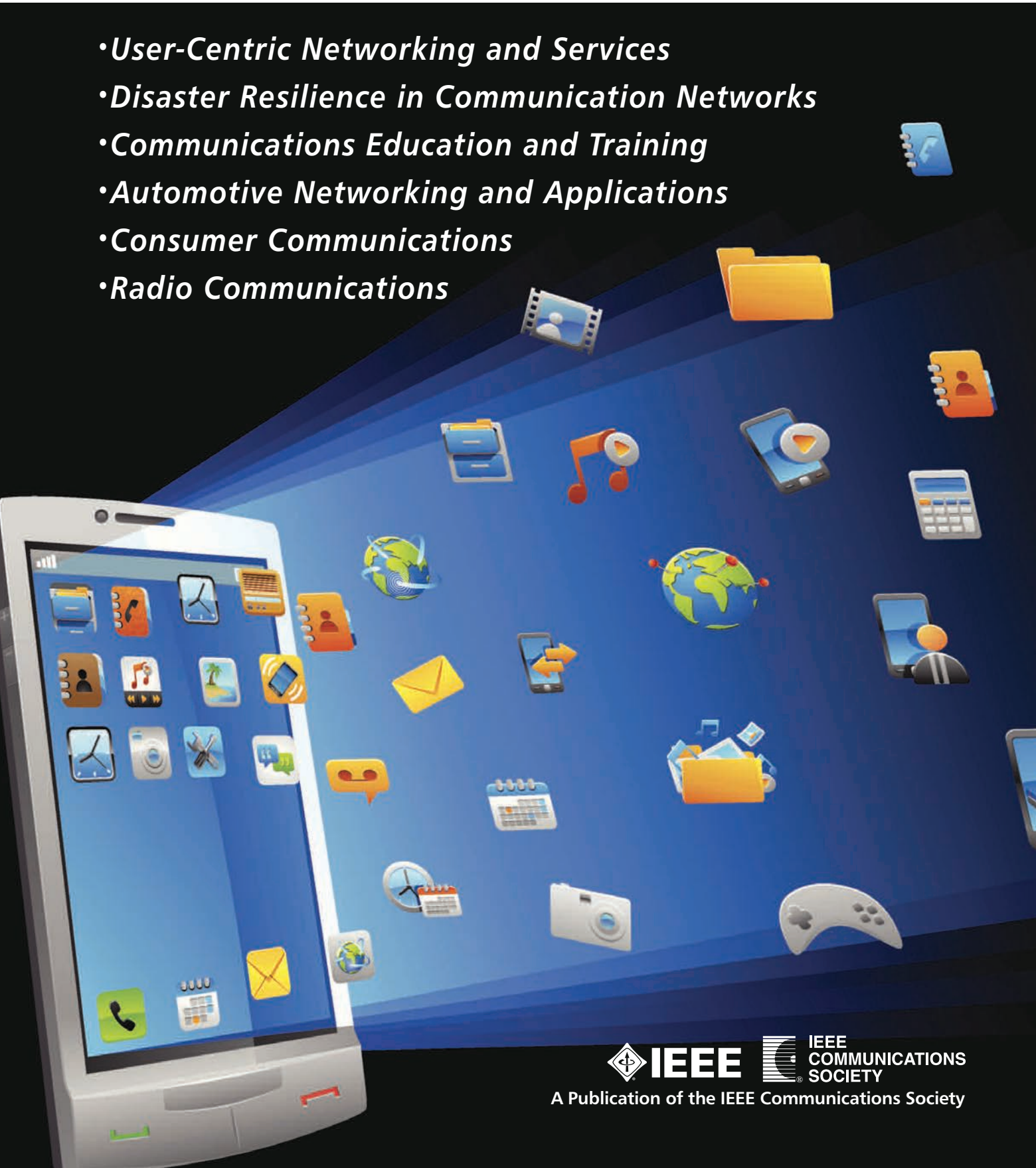
December 2014, Vol. 52, No. 12

IEEE Communications MAGAZINE

www.comsoc.org

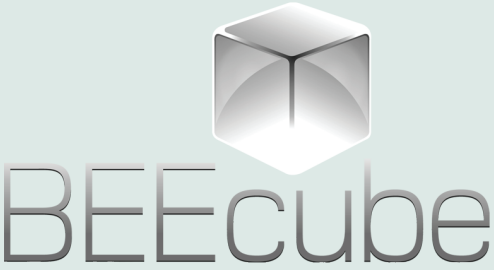
- *User-Centric Networking and Services*
- *Disaster Resilience in Communication Networks*
- *Communications Education and Training*
- *Automotive Networking and Applications*
- *Consumer Communications*
- *Radio Communications*



IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

IEEE Communications MAGAZINE

THANKS OUR CORPORATE SUPPORTERS

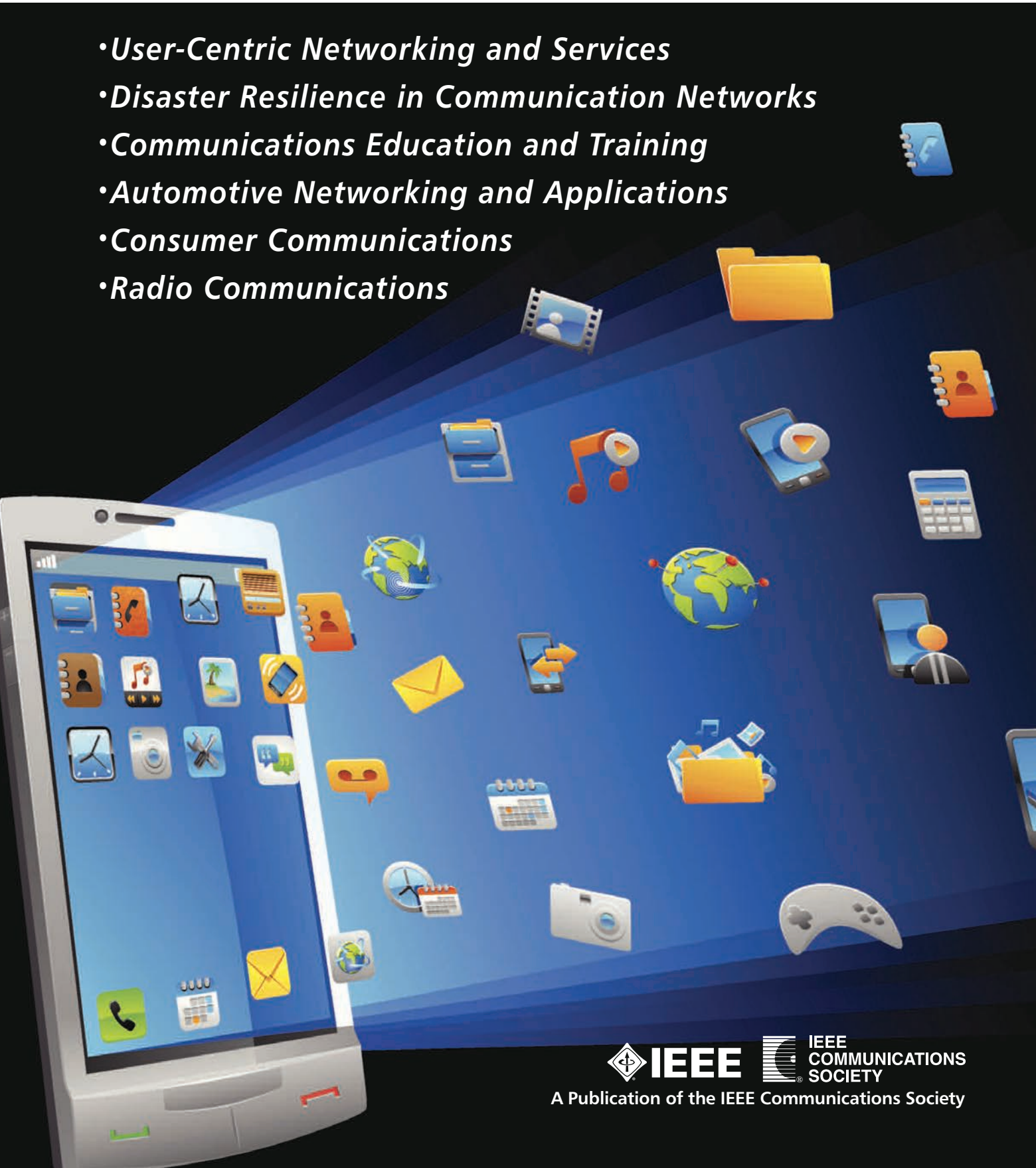


December 2014, Vol. 52, No. 12

IEEE Communications MAGAZINE

www.comsoc.org

- *User-Centric Networking and Services*
- *Disaster Resilience in Communication Networks*
- *Communications Education and Training*
- *Automotive Networking and Applications*
- *Consumer Communications*
- *Radio Communications*



IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

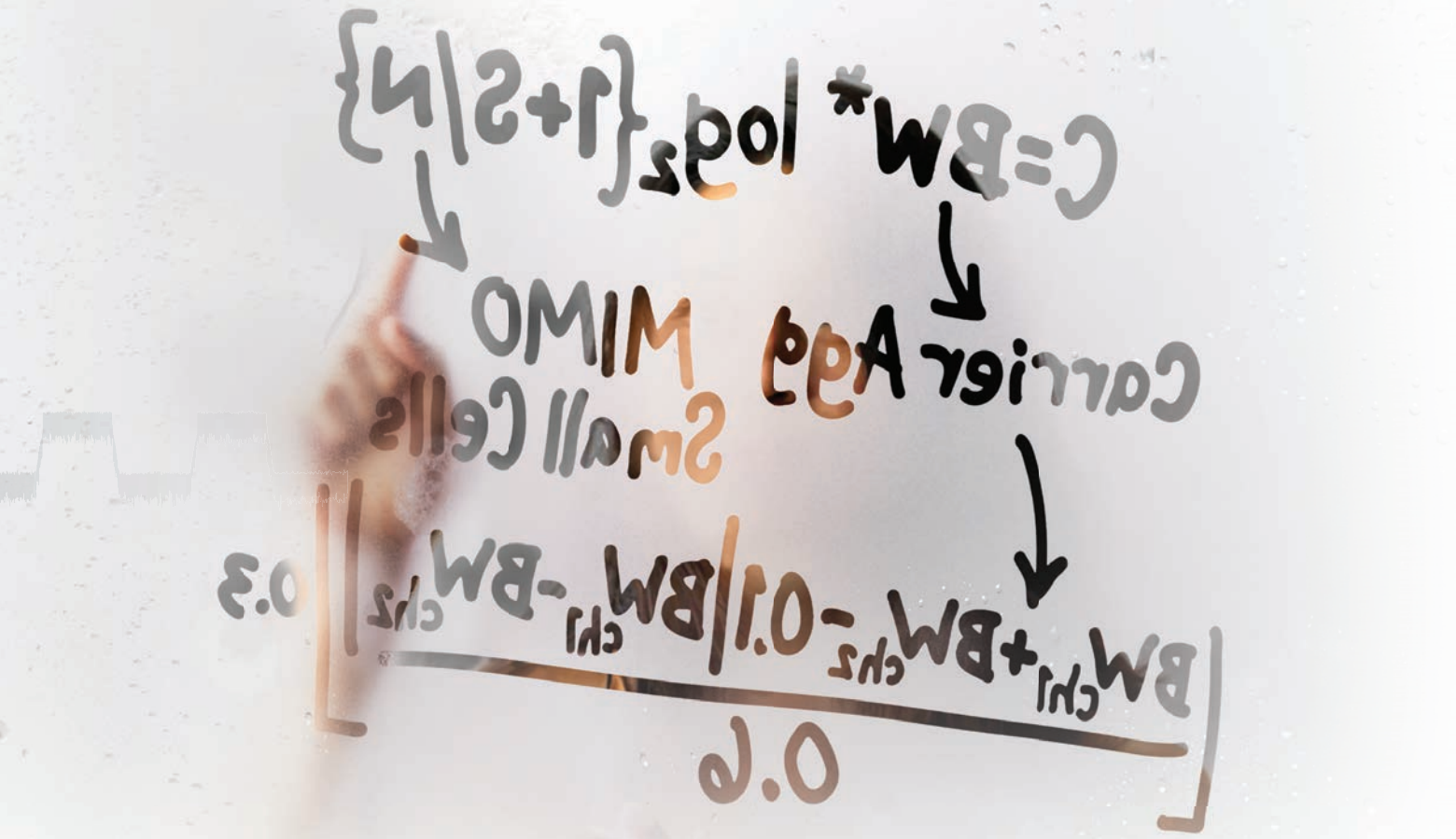
Eureka!

We'll help you get there.

Keysight 89600 VSA software

Insight. It comes upon you in a flash. And you know at once you have something special. At Keysight Technologies, we think precise measurements can act as a catalyst to breakthrough insight. That's why we offer the most advanced electronic measurement tools for LTE-A technology. We also offer sophisticated, future-friendly software. In addition, we can give you expert testing advice to help you design custom solutions for your particular needs.

HARDWARE + SOFTWARE + PEOPLE = LTE-A INSIGHTS



Download new LTE-A Technology and Test Challenge – 3GPP Releases 10,11,12 and Beyond
www.keysight.com/find/LTE-A-Insight

USA: 800 829 4444 CAN: 877 894 4414

Keysight W1715EP SystemVue MIMO channel builder

Keysight N9040B UXA signal analyzer
with 89600 VSA software

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

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

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

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

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

HARDWARE + SOFTWARE

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

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

PEOPLE

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

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

KEYSIGHT TECHNOLOGIES

Unlocking Measurement Insights

IEEE Communications Magazine

DECEMBER 2014, Vol. 52, No. 12

www.comsoc.org/commag

6 **The President's Page**
10 **Book Reviews**
11 **Global Communications Newsletter**
192 **Advertisers' Index**

User-Centric Networking and Services: Part 2

Guest Editors: Rute Sofia, Alessandro Bogliolo, Fikret Sivrikaya, Huiling Zhu, Olivier Marce, and David Valerdi

16 **Guest Editorial**

18 **Cooperative Relaying in User-Centric Networking under Interference Conditions**
Tauseef Jamal and Paulo Mendes

26 **Spontaneous Smartphone Networks as a User-Centric Solution for the Future Internet**
Gianluca Aloi, Marco Di Felice, Valeria Loscrì, Pasquale Pace, and Giuseppe Ruggeri

34 **Content Dissemination in Vehicular Social Networks: Taxonomy and User Satisfaction**
Farouk Mezghani, Riadh Dhaou, Michele Nogueira, and André-Luc Beylot

41 **A Trajectory-Based Recruitment Strategy of Social Sensors for Participatory Sensing**
Fei Hao, Mingjie Jiao, Geyong Min, and Laurence T. Yang

48 **Security and Performance Challenges for User-Centric Wireless Networking**
Pantelis A. Frangoudis and George C. Polyzos

Disaster Resilience in Communication Networks: Part 2

Guest Editors: Michele Nogueira, Piotr Chołda, Deep Medhi, and Robert Doverspike

56 **Guest Editorial**

58 **Network Adaptability to Disaster Disruptions by Exploiting Degraded-Service Tolerance**
S. Sedef Savas, M. Farhan Habib, Massimo Tornatore, Ferhat Dikbiyik, and Biswanath Mukherjee

66 **Enabling Disaster-Resilient 4G Mobile Communication Networks**
Karina Gomez, Leonardo Goratti, Tinku Rasheed, and Laurent Reynaud

74 **EmergeNet: Robust, Rapidly Deployable Cellular Networks**
Daniel Iland and Elizabeth M. Belding

81 **Exploiting the Use of Unmanned Aerial Vehicles to Provide Resilience in Wireless Sensor Networks**
Jó Ueyama, Heitor Freitas, Bruno S. Faiçal, Geraldo P. R. Filho, Pedro Fini, Gustavo Pessin, Pedro H. Gomes, and Leandro A. Villas

88 **Network Virtualization for Disaster Resilience of Cloud Services**
Isil Burcu Barla Harter, Dominic A. Schupke, Marco Hoffmann, and Georg Carle

free Tutorials Now™

ONLINE TUTORIAL
from IEEE Communications Society
www.comsoc.org/freetutorials

EFFICIENT USE OF SATELLITE RESOURCES SUPPORTING IP NETWORKS

Satellites provide IP-based converged voice, video and data communications access to deployed military forces. However, there are issues related to efficient use of expensive and limited space spectrum. There are overheads associated with the Internet Protocols (IP) and the need to encrypt critical messages. There are time delays in TCP-based flows. Rain fade can create increased bit errors, especially at higher frequencies. This seminar describes the latest technological innovations that address these issues. Adaptive coding/modulation increases effective capacity for channels susceptible to rain fade. Overlapped carriers enable uplink and downlink transmissions to share satellite bandwidth. More powerful satellites enable use of more efficient modulation and coding. TDM/TDMA/DAMA technology enables dispersed terminals to share spectrum taking advantage of statistical multiplexing in IP networks. The impact of IP overheads can be reduced, especially for voice traffic. Data traffic can be compressed. The performance of TCP-based transactions can be enhanced. A case history will illustrate the potential gains from combining these mitigations in a representative TDM/TDMA/DAMA satellite network.

LIMITED TIME ONLY AT >> WWW.COMSOC.ORG/FREETUTORIALS

IEEE COMMUNICATIONS SOCIETY

For this and other sponsor opportunities, please contact Susan E. Schneiderman, Business Development Manager. Phone: 732-562-3946. Email: ss.ieeemedia@ieee.org.

FREE ACCESS SPONSORED BY

Anritsu

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, Sean Moore, Centripetal Networks, CTO, 20 Mendelssohn Drive, Hollis, NH, USA 03049; tel: +1(603) 886-7343, e-mail: smoore-phd@ieee.org.

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 Osman Gebizlioglu, Associate Editor-in-Chief (Osman.Gebizlioglu@huawei.com). All submissions will be peer reviewed.

BPA WORLDWIDE

Communications Education and Training: Expanding The Student Experience

Guest Editors: David G. Michelson, Maria Trocan, and Wen Tong

90 **Guest Editorial**

98 **A Project Oriented Learning Experience for Teaching Electronics Fundamentals**
Frédéric Amiel, Dieudonné Abboud, and Maria Trocan

101 **Bringing an Engineering Lab into Social Sciences: Didactic Approach and an Experiential Evaluation**
Jesús Cano, Roberto Hernández, and Salvador Ros

Radio Communications: Components, Systems, and Networks

Series Editors: Thomas Alexander and Amitabh Mishra

108 **Series Editorial**

110 **MIMO for Millimeter-Wave Wireless Communications: Beamforming, Spatial Multiplexing, or Both?**
Shu Sun, Theodore S. Rappaport, Robert W. Heath, Jr., Andrew Nix, and Sundeep Rangan

122 **MIMO Precoding and Combining Solutions for Millimeter-Wave Systems**
Ahmed Alkhateeb, Jianhua Mo, Nuria González-Prelcic, and Robert W. Heath Jr.

132 **IEEE 802.11ad: Directional 60 GHz Communication for Multi-Gigabit-per-Second Wi-Fi**
Thomas Nitsche, Carlos Cordeiro, Adriana B. Flores, Edward W. Knightly, Eldad Perahia, and Joerg C. Widmer

Trends in Consumer Communications

Series Editors: Ali C. Begen, Mario Kolberg, and Madjib Merabti

142 **Series Editorial**

143 **Homenet3D: A New View on Home Network State**
Grenville Armitage and Dominic Allan

150 **The Greening of Spectrum Sensing: A Minority Game-Based Mechanism Design**
Mouna Elmachkour , Essaid Sabir, Abdellatif Kobbane, Jalel Ben-Othman, and Mohammed El koutbi

157 **Video Adaptation for Consumer Devices: Opportunities and Challenges Offered by New Standards**
James Nightingale, Qi Wang and Christos Grecos, and Sergio Goma

Automotive Networking and Applications

Series Editors: Wai Chen, Luca Delgrossi, Timo Kosch, and Tadao Saito

164 **Series Editorial**

166 **Cooperative Intelligent Transport Systems Standards in Europe**
Andreas Festag

173 **Implementing Virtual Traffic Lights with Partial Penetration: A Game-Theoretic Approach**
Ozan K. Tonguz, Wantanee Viriyasitavat, and Juan M. Roldan

183 **Intra-Car Multihop Wireless Sensor Networking: A Case Study**
Morteza Hashemi, Wei Si, Moshe Laifenfeld, David Starobinski, and Ari Trachtenberg

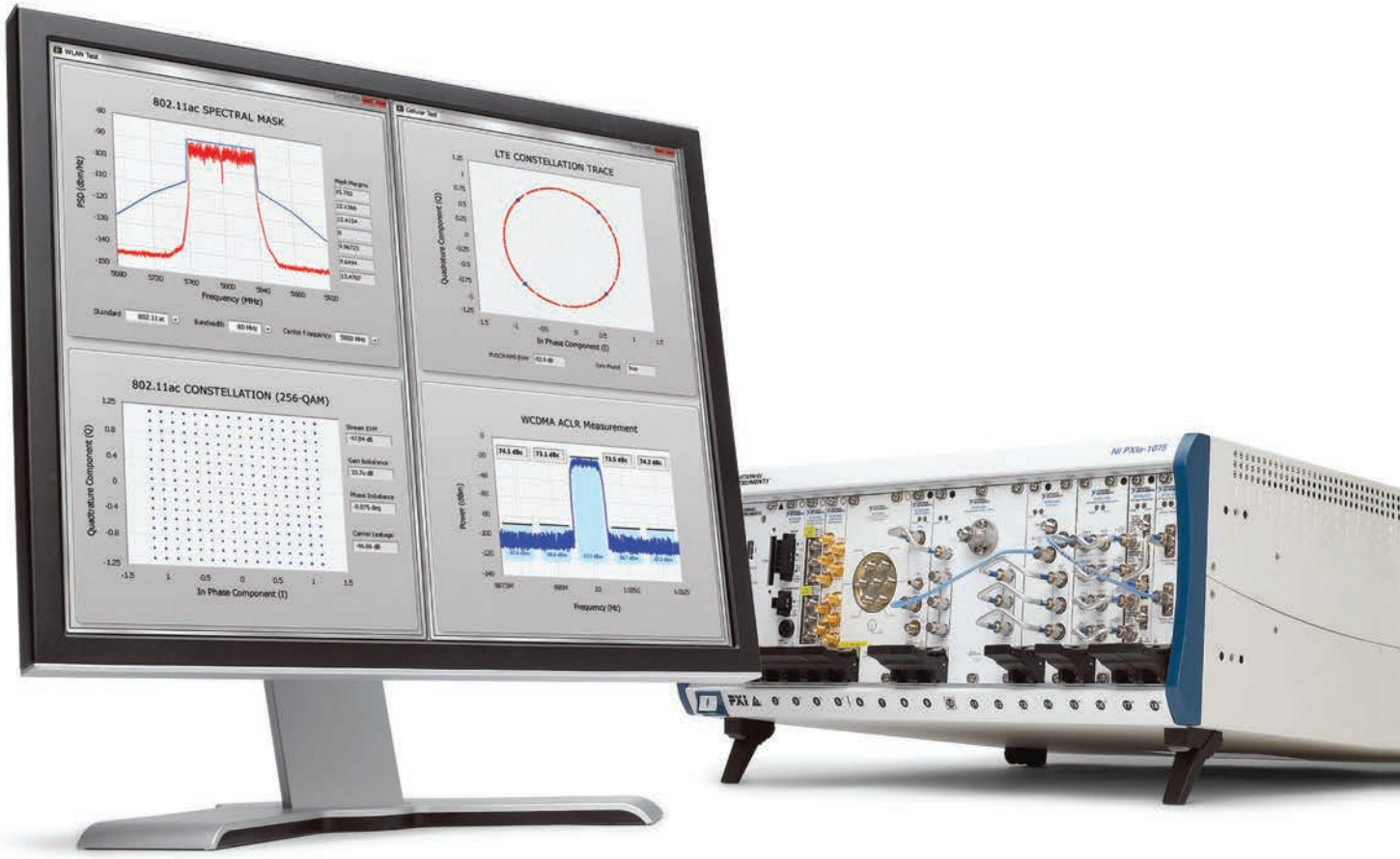
Currently Scheduled Topics

Topic	Issue Date	Manuscript Due Date
Communications Education and Training	May 2015	January 1, 2015
Internet of Things/M2M from Research to Standards	August 2015	January 15, 2015
Software Defined 5G Networks for Anything as a Service	September 2015	January 15, 2015

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

Redefining RF and Microwave Instrumentation

with open software and modular hardware



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

WIRELESS TECHNOLOGIES

National Instruments supports a broad range of wireless standards including:

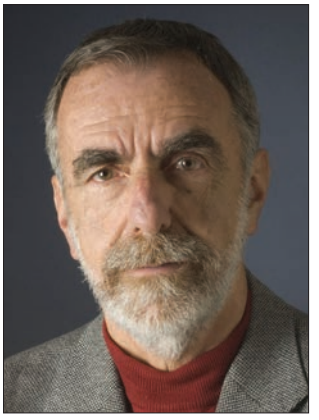
802.11a/b/g/n/ac	LTE
CDMA2000/EV-DO	GSM/EDGE
WCDMA/HSPA/HSPA+	Bluetooth

>> Learn more at ni.com/redefine

800 813 5078

NATIONAL INSTRUMENTS

HONORING OUR COLLEAGUES: COMSOC AWARDS



SERGIO BENEDETTO

LAJOS HANZO

The December page is devoted to IEEE Communications Society Awards and the Awards Committee that receives the nominations and processes them to identify the recipients. ComSoc Awards are meant to honor colleagues who in some ways, either via scientific/technical contributions or exemplary services, have reached significant, widely recognized achievements in our telecommunications community. Owing to their high significance and value, they need to be awarded through a fair and transparent process, and this is the essential, yet difficult and heavy task of the Awards Committee. It is my pleasure to introduce Lajos Hanzo, the Chair of the IEEE Communications Society Awards Committee, who will describe the awards and the procedure followed by the Awards Committee.

Lajos received his five-year Dipl-Ing./Master degree in electronics in 1976 and his doctorate in 1983. In 2009 he was awarded the honorary doctorate "Doctor Honoris Causa'' by the Technical University of Budapest. During his 38-year career in telecommunications he has held various research and academic posts in Hungary, Germany, and the UK. Since 1986 he has been with the School of Electronics and Computer Science, University of Southampton, UK, where he holds the chair in telecommunications. He has successfully supervised about 100 Ph.D. students, co-authored 20 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10 000 pages, published 1400+ research entries at IEEE Xplore, acted both as TPC and General Chair of numerous IEEE conferences, presented keynote lectures, and has been awarded a number of distinctions. Currently he is directing an academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) UK, the European Research Council's Advanced Fellow Grant, and the Royal Society's Wolfson Research Merit Award. He is an enthusiastic supporter of industrial-academic liaison and a Governor of the IEEE VTS. During 2008–2012 he was the Editor-in-Chief of the IEEE Press and a Chaired Professor also at Tsinghua University, Beijing. His research is funded by the European Research Council's Senior Research Fellow Grant. For further information on research in progress and associated publications please refer to http://www-mobile. ecs.soton.ac.uk

ACKNOWLEDGEMENTS

Over the years numerous valued colleagues have volunteered their precious time to serve in diverse capacities in honoring our deserving colleagues for their expectional contributions. The seamless operation of the process critically depends on the altruistic efforts of the nominators, of the colleagues writing support letters, of the distinguished colleagues serving on the Awards Committee, and on the tireless assistance of Carol Cronin, Carole Swaim, and Natasha Simonovski of ComSoc, just to mention a few.

The current Awards Committee members serving throughout the period of 2014-2016 are: Anja Klein, Geoffrey Y Li, Hussein Mouftah, Anthony Soong, Branka Vucetic; 2013-2015: Aria Nosratinia, Rong Pan, Ljiljana Trajkovic; 2012-2014: Costas Georghiades, Vincent Lau, Ian Oppermann, Jack Winters. Sincere gratitude for your staunch support! Last, but not least, the past chairs of the Awards Committee — most recently Andy Molisch, Vince Poor, and Roberto de Marca — also deserve tremendous credit for their selfless dedication, along with our Vice-President, Khaled Ben Letaif. This is our chance to pay tribute to all of them for the development of the ComSoc Awards Program.

TIME-SCALES

The ComSoc awards may be classified as paper awards, career awards, and service awards, which are presented at the ComSoc 'flagship' conferences ICC and Globecom. No doubt most of you Dear Readers would have participated in the Awards Luncheon Ceremonies, which are skillfully organized by Carole Swaim in collaboration with the Awards Committee Chair. Since ICC is usually in June, the selection of the best paper awards has to be concluded about six to eight weeks before ICC for presentation there, while the career and service awards are typically bestowed upon the winners at Globecom in December. In order to provide sufficient time for the Awards Committee to evaluate the nominations, the best paper nominations are typically due by the 15th of February, while the service and career awards must be submitted by the 1st of September.

THE AWARDS

The entire history of telecommunications may be traced back by simply observing the list of Edwin Howard Armstrong Achievement Award winners, which was first bestowed in 1958 (http://www.comsoc.org/about/memberprograms/comsoc-awards/armstrong). The entire spectrum of awards and winners can be followed at: http://www.comsoc.org/about/memberprograms/comsoc-awards/. Many of these prestigious awards were named after our distinguished predecessors, who shaped the evolution of telecommunications. We are also pleased to announce the recent approval of two new awards, namely a career award for a Distinguished Educator and a best paper award for a scientist younger than 30 years of age.

COMSOC BEST PAPER AWARDS

Most of the ComSoc journals have their dedicated awards, albeit there are some exceptions as well, especially when it comes to the journals and awards that are shared by several societies. To elaborate a little further, below is the current list of awards.

The *Best Tutorial Paper Award* is bestowed upon the authors of an outstanding tutorial paper published in any Communications Society magazine or journal in the past five calendar years.

The *Charles Kao Award for Best Optical Communications & Networking Paper* will be given to the distinguished authors of papers published in the *OSA/IEEE Journal on Optical Communications & Networking (JOCN)* that open new lines of research, envision bold approaches to optical communication and networking, formulate new problems to solve, and essentially extend the field of optical communications and networking. Papers published in the prior three calendar years of JOCN are eligible for the award.

The *Fred W. Ellersick Prize* honors the authors of an excellent paper published in any Communications Society magazine in the past three calendar years.

The *Heinrich Hertz Award for Best Communications or Wireless Communications Letter* is given to the authors of an excellent correspondence/letter in these publications during the previoous three calendar years.

The *Leonard G. Abraham Prize in the Field of Communications Systems* is dedicated to an exceptional paper published in the *IEEE Journal on Selected Areas in Communications* during the past three calendar years.

The *Stephen O. Rice Prize in the Field of Communications Theory* honors the authors of what was deemed to be the best paper published in the *IEEE Transactions on Communications* during the past three calendar years.

The *William R. Bennett Prize in the Field of Communications Networking* distinguishes the authors of the best paper published in the *IEEE/ACM Transactions on Networking* during the past three calendar years.

The *IEEE Communications Society Award for Advances in Communication* is awarded for an outstanding paper in any Communications Society publication on a promising new subject in the preceding 15 calendar years.

The *ComSoc & Information Theory Joint Paper Award* honors the authors of a paper that is relevant to both of the above-mentioned Societies and was published in any of the Communications Society or the Information Theory Society journals within the past three years. This joint award is judged by both ComSoc and the Information Theory Society, and the joint committee is chaired in alternate years by the chairs of the two Awards Committees.

The *IEEE Marconi Prize Paper Award in Wireless Communications* is a joint award with the IEEE Signal Processing Society, and it is dedicated to a paper published in the *IEEE Transactions on Wireless Communications*.

The *IEEE Communications Society Young Author Best Paper Award* honors the author(s) of an especially meritorious paper dealing with a subject related to the technical scope of the Society and who, upon the date of submission of the paper, is less than 30 years of age.

It is worth mentioning that in 2011 the period of eligibility for the majority of the above-mentioned awards was increased to three years, so that the medium-term impact of the papers was more quantifiable. As an exception, the time period of the *Best Tutorial Paper Award* was extended to a five-year eligibility window, while that of the *IEEE Communications Society Award for Advances in Communication* is as long as the preceding 15 calendar years.

ComSoc Career Awards

The *Edwin Howard Armstrong Achievement Award* is bestowed in recognition of outstanding contributions over a period of years in the field of communications technology, which was awarded to highly acclaimed leaders of our field since 1958, including Claude Shannon in 1973. Just to reminisce for a moment about their contributions over the past decade: 2014 – Andrea Goldsmith; 2013 – John M. Cioffi; 2012 – Donald Cox; 2011 – no award; 2010 – no award; 2009 – H. Vincent Poor; 2008 – Sergio Benedetto; 2007 – Norman C. Beaulieu; 2006 – Larry J. Greenstein; 2005 – no award; 2004 – Hussein Mouftah.

The *Distinguished Industry Leader Award* is bestowed upon an executive leader, who was able to inspire substantial advances and open up new research directions in the information and communications business area.

The *Industrial Innovation Award* is given to colleagues for their major industrial accomplishments, standards, deployment of important processes or products, etc., that are of substantial benefit to the public in the field of communications and information technologies. The innovations must be visible beyond the company or institution where the contribution was made.

The *Award for Public Service in the Field of Telecommunications* is given to an individual for major contributions to the public welfare through work in the field of telecommunications.

The *IEEE Communications Society Education Award* is bestowed upon a Communications Society member, who is an acclaimed educator and who has made distinguished contributions to education within the field of ComSoc.

ComSoc Service Awards

The *Donald W. McClellan Meritorious Service Award* is bestowed upon an individual for outstanding long-term service and leadership in the welfare of the IEEE Communications Society. The recipient must be a ComSoc member.

The *Harold Sobol Award for Exemplary Service to Meetings & Conferences* is given to a colleague for exemplary service to IEEE Communications Society meetings and conferences over a sustained period of time.

The *Joseph LoCicero Award for Exemplary Service to Publications* is given to a colleague for his/her dedicated service to IEEE Communications Society publications over a sustained period of time.

The *ComSoc/KICS Exemplary Global Service Award* is for fostering successful partnerships between ComSoc and its sister societies, as well as for encouraging a spirit of mutual support and respect in the international communications community. This award is particularly distinguished, since it is co-sponsored by ComSoc and the Korea Information and Communications Society (KICS), our sister society located in Korea. This award is presented jointly by the presidents of both these two societies.

Work of the Awards Committee

The success of the Awards Committee's work critically hinges on the community-at-large both in terms of nominating excellent papers and meritorious candidates for the awards.

Best Paper Nomination Process: Apart from the usual nominations from members of the research community, the Chair communicates with the EICs of all eligible journals to ensure that the most deserving papers were never overlooked. The EICs have had their own realiable processes for identifying the best papers to be nominated. In some cases the EICs forwarded the papers having the highest number of citations to their respective steering committees, and their members were also involved in pre-ranking the papers, which were then nominated by the EIC to the Awards Committee.

An exceptionally high-integrity process was used in all cases, as exemplified by the process used by Prof. Schober, the EIC of TCOM. He invited nominations from the Editors and also screened the nominations made through Manuscript Central by Editors at the time of accepting each paper. In addition, he looked up the most highly cited TCOM papers in the relevant period. Based on this, he had an initial list of nine papers. He then formed an *ad hoc* committee consisting of Des Taylor, Michele Zorzi, Ender Ayanoglu, and Alberto Zanella (Senior Editors of TCOM). This committee felt that three of the nine papers from last year's round should also be nominated again. For the remaining six papers, the EIC asked each committee member to evaluate one or two papers in terms of their suitability for an award. Based on these evaluations, this committee identified two additional papers that they felt were worthy candidates for nominations to the Awards Committee.

Best Paper Nomination Statistics: In excess of 50 papers were nominated, but there were numerous papers that were nominated more than once. For these papers the highest-quality nomination was retained for the committee's consideration. A total of 47 valid nominations were received by the 15th of February deadline and in excess of 150 paper evaluations were provided by the 12 Awards Committee members.

Best Paper Selection: The Chair formed a sub-committee for each award and the members were asked to evaluate the novelty, readability, rigor, the number of citations, and any other aspects such as the quality of references used, the quality of illustrations, conclusions, etc.

This process resulted in the three top candidates for each award, which were then finally voted on by the entire committee in the ensuing Phase 2. A similar twin-phase process was also used for deciding upon the Career Awards as well as Service Awards, and in all cases there was an almost unanimous agreement concerning the final award winners. The chair did not vote and there was no need for him to break any ties.

Our hope is that you Dear Colleagues have found this brief article to be informative and that we might have suceeded in stimulating you to volunteer some of your precious time to nominate a truly exceptional paper or individual!

OMBUDSMAN

COMSOC BYLAWS ARTICLE 3.8.10

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

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

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

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

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

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

Technological highlights: power measurement

- Unrivaled range of easy-to-use USB sensors
- Highest precision and measurement speed
- Wideband sensors up to 44 GHz
- Thermal sensors up to 110 GHz
- Time domain measurement of radar pulses
- Ultra-fast statistical analyses

ROHDE&SCHWARZ

IPV6 DEPLOYMENT AND MANAGEMENT

BY MICHAEL DOOLEY AND TIMOTHY ROONEY

WILEY/IEEE PRESS, 2013, ISBN 978-1-118-38720-7, SOFTCOVER, 204 PAGES

REVIEWER: KRZYSZTOF LOZIAK

This book is a definitive quick guide to the IPv6 World, suitable for readers who would like to become familiar with IPv6, but do not have enough time to go through many different books or dig into RFCs. Dooley and Rooney cover the topic essentials such as: the IPv6 address composition, types of IPv6 addresses (global, local, multicast), and, of course, the IPv6 packet header structure. Unicast, anycast, and multicast communication types are described. The IPv6 address space and allocation schemes are investigated, allowing the reader to understand the enormous number of IPv6 addresses in comparison to the previous IP version, as well as the globe-wide address saturation problem. The authors describe not only the IPv6 protocol, but also the whole family of protocols operating around it: NDP, SEND, MRD, ICMPv6, DAD, etc. Additionally, MobileIPv6 is covered, emphasizing the fact that IPv6-enabled nodes can communicate seamlessly while moving from site to site or link to link, taking benefits of different access technologies, such as 4G, WiFi, etc.

A huge part of this book focuses on IPv6 and IPv4 coexistence technologies, such as the dual stack, tunneling techniques, and translation approaches, given with several examples of the solutions preferred by service providers. Coexistence is the most important milestone on the roadmap toward the full IPv6 implementation. Therefore, better understanding of possible implantation mechanisms allows service providers to choose a preferable strategy and plan their IPv6 deployment process.

The IPv6 address planning process precedes the deployment phase. Starting from Internet Registries, Regional Internet Registrars, and then National and Local ones, the process is described with examples of useful algorithms, for example: the Best-fit Method, Sparse Allocation, Random Allocation, or DHCPv6 Prefix Delegation. Some selected strategies to deal with the obtained IPv6 addressing block for ISPs are presented and extended with valuable examples.

Finally, a reader can learn how to face the problem of managing the IPv4/IPv6 network when 'the network' still remains physically 'the same network', but we need to know which steps should be taken into consideration to add, remove, or reallocate devices or subnets.

Summarizing, I would like to truly recommend this book to anyone interested in the technology: IPv6 newbies, enthusiasts or skeptics, as well as network engineers, administrators searching for IPv4 to IPv6 migration motivation, students willing to touch a topic with a higher level overview, and anyone who wants to be prepared for what is inevitable in the world of IP networking.

FLOW NETWORKS: ANALYSIS, AND OPTIMIZATION OF REPAIRABLE FLOW NETWORKS, NETWORKS WITH DISTURBED FLOWS, STATIC FLOW NETWORKS AND RELIABILITY NETWORKS

BY MICHAEL T. TODINOV

ELSEVIER, 2013, ISBN 978-0-12-398396-1, HARDCOVER, 247 PAGES

REVIEWER: JACEK RAK

There are many books available on the market that address the issues of design and optimization of static flow networks. However, even though applications of such networks are numerous, the idea of static flow networks seems not to respond properly to flow disruptions in real networks being results of faults of network elements, causing unpredictable flow fluctuations.

The book by Michael T. Todinov is intended to provide a comprehensive state of the art discussion of design and optimization methods of flow networks. The added value of this book is that, unlike many other positions, the main focus is on design and optimization methods of networks with disturbed flows (which is the common scenario for almost all real networks), requiring implementation of recovery scenarios (e.g. redirection of affected flows onto alternate paths after failures). Frequency of failures, as well as the extent of possible losses, make fault recovery an indispensable element of real network design and optimization processes.

The book consists of 14 chapters, all providing an in-depth analysis of the related aspects. The first four chapters refer to basic aspects of design and optimization of static flow networks, as they are necessary in understanding the theory of repairable flow networks and networks with disturbed flows. In particular, the book includes presentations of novel algorithms for flow maximization, and provides proofs of important theorems, such as a new fundamental dual network theorem for static networks.

Discussions on networks with disturbed flows start from Chapter 5, and include presentations of fast augmentation algorithms aimed at restoring the maximum possible throughput flow in a network after an edge failure. Chapters 7 and 8 are the next important parts highlighting the aspects of reliability of the throughput flow and reliability networks. In Chapter 9 another important result of the author is presented. It refers to the average production availability of repairable flow networks under the assumption of independent work of network edges characterized by the exponential distribution of times to failure. Then Chapter 10 focuses on topological characteristics influencing the performance of repairable flow networks. Conclusions following from this chapter contending that any two networks composed of even identical numbers of components may have visibly different performance characteristics, are next utilized in Chapter 11 to propose the topology optimization method of repairable flow networks and reliability networks. The last three chapters of the book refer to special cases of repairable networks with merging flows, flow optimization in repairable flow networks, as well as failure acceleration stresses.

Although the book is not related directly to computer communications aspects (solutions presented here are valid for many networked systems, including, for instance, oil and gas production systems), and requires the reader to have some knowledge/mathematical background on network optimization issues, as well as formal mathematical reasoning, it is undoubtedly a solid publication targeted at graduate students and academia, as well as industry researchers. The readers should find this book useful mainly due to the presentation of models and algorithms applicable to real network problems.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

December 2014
ISSN 2374-1082

MEMBER RELATIONS

North America Region

Interview with Merrily W. Hartmann, Director of the North America Region

By Stefano Bregni, Vice-President for Member Relations, and Merrily W. Hartmann, Director of the North America Region

This is the fourth article in the series of eight, opened in September and published monthly in the *Global Communications Newsletter*, which covers all areas of IEEE ComSoc Member Relations. In this series of articles, I introduce the seven Member Relations Directors (namely: Sister and Related Societies; Membership Programs Development; AP, NA, LA, EAME Regions; Marketing and Industry Relations) and the Chair of the Women in Communications Engineering (WICE) Standing Committee. In each article, one by one they present their activities and plans.

Stefano Bregni

Merrily W. Hartmann

In this issue, I interview Merrily W. Hartmann, Director of the North America Region.

Merrily received her B.S. degree in Mathematics and Computer Sciences from the University of Illinois, Chicago. She retired from SBC Communications, Inc. (formerly Southwestern Bell and now AT&T) being Executive Director of Global Markets Sales Support, managing all sales operations for SBC's 200 largest customers, in 2000 after 25 years of service. Prior to joining Southwestern Bell, she started her career with Bell Telephone Laboratories (Naperville, IL).

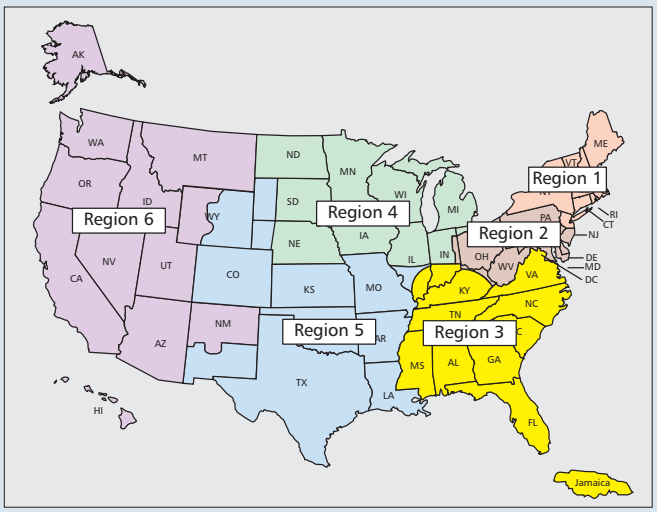
Merrily has been a member of IEEE for 31 years. Currently, she is serving on the ComSoc Board of Governors as a Member-at-Large and as the North America Region Director. Previously, she served as Director of Conference Operations (2008-2011, 2013) and Member-at-Large of the GIMS Committee (2006-2007), which is responsible for providing strategic guidance and management oversight of the Society's flagship conferences, ICC and GLOBECOM.

It is my pleasure to interview her and to have this opportunity to focus on the North America Region, to present how it is organized and how it operates.

Begni: *Hello Merrily. Let us begin by presenting how the North America Region is organized geographically.*

Hartmann: The IEEE Member and Geographic Activities (MGA) organization is divided into 10 regions. In turn, ComSoc manages its member activities via four regions, which are composed of the IEEE regions as shown in the accompanying table.

IEEE	ComSoc
Regions 1-6 (United States)	North America Region
Region 7 (Canada)	
Region 8 (Europe-Middle East-Africa)	EMEA Region
Region 9 (Latin America)	LA Region
Region 10 (Asia Pacific)	AP Region



Bregni: Might you give us some membership statistics for Regions 1-7?

Hartmann: At the end of 2013, IEEE had 431,191 members. Out of them, 221,410 (51%) were in Regions 1-7, that is ComSoc's NA Region. Among them, 20,525 were ComSoc members. In the ComSoc NA Region, there are 92 Chapters, ranging in size from 2,000 to 5,000 members. Each of the seven IEEE Regions includes from 8 to 16 ComSoc Chapters.

Bregni: So, what is the mission of all of those ComSoc Chapters? For what purpose have they been created?

Hartmann: IEEE MGA defines a Chapter as a "technical subunit of one or more Sections." Specifically, ComSoc Chapters provide a local connection for our society members to interact and engage regarding ComSoc's field of interest. It is ComSoc's challenge to create and conduct activities that will enable these interactions/engagements.

Bregni: That is, in practice? Could you provide some actual examples of activities organized by ComSoc Chapters in North America?

Hartmann: Chapter activities include the following: Distinguished Lecturer talks, social events, member status elevation sessions, workshops, seminars, special events, etc. Coming up with interesting and relevant programming throughout the year requires a lot of creativity, planning, and coordination by our Chapter Chairs and their committees. To provide some specific examples, I'd like to highlight the special events held by our New Jersey Coast ComSoc Chapter in 2013, which contributed to their receiving the 2014 ComSoc Chapter Achievement Award for the NA region:

•The chapter proposed honoring ComSoc's 60th anniversary

(Continued on Newsletter page 5)

CONFERENCE REPORT

1st FUSECO (FUture SEamless COmmunications) Forum Asia (FFAsia 2014)

By Thomas Magedanz, Florian Schreiner, and Anne Halbich, Fraunhofer FOKUS, Germany; I. Narayana and Fajar Nugroho, TELKOM Indonesia; Thomas Michael Bohnert, Zurich University of Applied Sciences, Switzerland

The 1st FUSECO (FUture SEamless COmmunications) Forum Asia (FFAsia 2014), organized jointly by PT TELKOM Indonesia, and Fraunhofer FOKUS, was held in Bali, Indonesia from June 9-10, 2014. With around 300 attendees from 14 countries, this event proved to be a real premier international forum in Asia, discussing different technical and business aspects of emerging ecosystems within Smart Cities and beyond.

Under the technical lead of Prof. Dr. Thomas Magedanz, an senior IEEE member for more than 20 years, and Ir. Joddy Hernady, MSEE, Group Head of Innovation and Design Center, Telkom Group Indonesia, this two day event featured four keynotes, six technical sessions, one vendor session, and a final panel discussion on the challenges and opportunities in the establishment of Smart Cities infrastructures. In addition, vendor exhibitions from Huawei, Fiberhome, ZTE, and CISCO, including Fraunhofer, were presented. Attendees had many opportunities to learn about the state of the art in Smart City enabling technologies and international best practices.

FFAsia Day One

After the opening by Prof. Dr. Thomas Magedanz on behalf of Fraunhofer FOKUS and Ir. Joddy Hernady, Telkom Group Indonesia, and Ir. Indra Utoyo, MSc, Chief of Innovation and Strategy Officer, Telkom Group Indonesia on behalf of Arief Yahya, the CEO of PT TELKOM Indonesia, two keynotes were held.

Prof. Dr. Radu Popescu-Zeletin presented in his keynote "Information & Communication Technology Convergence Enabling Smart Cities," the technological pillars of Smart Cities ICT platforms and stressed the important role of (big) data and the various required business processes. In the second keynote, "Enabling a Converged World Through Ecosystem Solution," Indra Utoyo presented the Indonesian Digital Network initiative from PT Telkom. In addition, he also discussed TELKOM's ecosystem framework featuring a business incubator and an experience center (Digital Lounge and Loop Station).

The first session entitled "Digital Lifestyle and Smart City Applications as Drivers for Mobile Broadband Network Evolution," chaired by Prof. Alfonso Ehijo from the Univ. de Chile, featured four talks from Asia that set the floor for the following Smart City considerations, namely addressing major digital lifestyle changes due to powerful new multimedia devices and describing potential Smart City application domains.

The second session on "Smart City Network – Evolution Path from LTE towards 5G and SDN," chaired by Prof. Rui Aguiar from the Universidade de Aveiro, Instituto de Telecomunicacoes, Portugal and Prof. Magedanz featured three talks on the requirements for emerging 5G and mobile broadband networks gathered by the NGMN Alliance, and the role of emerging SDN technologies for the flexible implementation of different virtual services in Japan, including virtualized IMS services, disaster tolerant services, and SDN-based Smart City services in Korea.

The third session, "Internet of Things/M2M as Backbone for Smart Cities," was chaired by Dr. Adel Al-Hezmi, from Fraunhofer FOKUS, Germany, and Prof. Noel Crespi, from Telcom Sued Paris, France. Based on an overview of SIM and NO SIM based IoT/M2M markets, an overview of the OneM2M Alliance standards was given. After that Telkomsel and Entel presented, from an operator perspective, interesting aspects on APIs provisioning for developers as well as data analytics, respectively.

Impressions of the FUSECO Forum 2014.

At a Balinese beach party and buffet the delegates had the oppotunity to relax and enjoy local dances, music, and food, while discussing questions and making new friends.

FFAsia Day Two

The second day started with two keynotes. Dr. Roberto Minerva from Telecom Italia talked about "Mastering the Innovation Challenges of the Future Network Operators in an Emerging IoT World," and stressed the major challenge of network operators in becoming successful Internet actors. He highlighted the potential of IoT for future telco business, ending with a new IEEE IoT initiative. Ir. Rizkan Chandra, M.Sc highlighted the challenges of the transition from a telecommunications company, like TELKOM Indonesia, into a digital company. He mentioned that it requires a transformation on People, Processes and Technology. In the case of TELKOM, changing the mindset and finding new business models were the biggest challenges.

Session 4, "Future Internet Technologies and Enablers as Foundation for Smart Cities," chaired by Serge Fdida from the UPMC Sorbonne University, France, featured two talks giving the European perspective of Smart City platform and application facilitation by means of a future Internet software toolbox (FIWARE), plus the related Software Development Kit (FILAB) and a corresponding validation use case for the eHealth domain given by the European FISTAR project.

Session 5 and Session 6 looked at "Smart Cities: Best Practices and Current Roleout Plans." Session 5 was chaired by Dr. Niklas Blum from Fraunhofer FOKUS, Germany and Prof. Akihiro Nakao from Tokyo University, Japan.

The first talk provided an initial market overview of Asian Smart City initiatives, stressing the importance of a "thinking platform" rather than a loose collection of individual applications. The second talk featured the "Ganesha Maturity Model," the approach adopted by ITB for the Indonesian Market. The last talk illustrated the steps adopted in South Africa in the context of Smart Energy.

Session 6 was chaired by Prof. Alfonso Ehijo from the Universidad de Chile and Assist. Prof. Dr. Supavadee Aramvith from Chulalongkorn University, Thailand. The first talk showed the initial steps toward Smart City Emergency services implementation taken in Thailand; the second talk showed the Smart Objects approach from France; the third talk showed some advances in surveillance and emergency infrastructures in Japan as well as the need for application driven SDNs. Finally, Orange highlighted the experiences gained in Smart City M2M services.

A dedicated Best Practices Vendor Session brought Huawei, Fiberhome, and ZTE to the stage, presenting their current engagements in cloud-based IMS deployments, Smart City portals and solutions in China.

The final panel, uniting all the main speakers, discussed interactively with the audience the challenges and opportunities in the establishment of Smart Cities ecosystems and infrastructures. It has become clear that the business cases for Smart Cities are

(Continued on Newsletter page 4)

CONFERENCE REPORT

Highlights from the 19th European Conference on Networks and Optical Communications (NOC 2014)

By Guido Maier, NOC2014 TP CoChair, Politecnico di Milano, Italy

IEEE ComSoc has technically co-sponsored the 19th European Conference on Network and Optical Communications, held in Milano, Italy, on June 4-6, 2014 (http://www.noc-conference.com).

The European Conference on Network and Optical Communications (NOC) was originally started in 1986 and has been run on a yearly basis in Europe. Up to last year's edition, it used to combine the Conference on Optical Cabling and Infrastructure (OC&I) within a single even, but starting in 2014 it will focus solely on NOC.

The goal of the conference is to present high-quality results in the field, and to provide a framework for research collaboration through focused discussions that will designate future research efforts and directions, as well as a forum for the promotion of new opportunities from industry, institutes of technology, research centers, and academia. Despite being naturally more oriented to the European research community on optical networks (both academic and industrial), NOC has so far attracted many attendees and presenters from non-European countries, especially the USA, China, and India, as well as South-America. NOC is a well consolidated European conference with international stature and broad coverage of topics related to optical networking and communications. Most recent editions have been in Spain (Polytechnic University of Catalunya in 2012) and Austria (Technical University of Graz in 2013).

The 19th European Conference on Network and Optical Communications (NOC 2014) was held in Milan, from June 4 to 6, 2014. The host university was Politecnico di Milano, one of the most prestigious technical education institutions in the country, celebrating its 150th anniversary in 2013. Politecnico offered its main campus "Leonardo" (Milano Città Studi) as the venue for the conference. The General chair, Prof. Achille Pattavina, and most of the organizing committee were from the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB) of Politecnico di Milano.

Through its technical program, the conference has provided an opportunity for the academic and industrial communities to meet and address new research challenges, share solutions, and discuss issues on optical networking and photonic equipment technology. NOC has addressed research topics in the optical networks area, such as core, metro, and access network planning design and modeling, optical communications (both fiber and free-space), and photonic devices. New topics such as optical mobile backhauling, optical interconnects, intra-datacenter networking, and software-defined optical networks have also been discussed.

NOC 2014 conference banquet in the ballroom of "Conte Biancamano" trans-Atlantic liner, inside the National Science and Technology Museum Leonardo da Vinci in Milan, June 5th, 2014.

Workshop "Optical Access Network Virtualization: from Unbundling to Open Access", co-organized by NOC 2014 and the Italian national research project ROAD-NGN, room "Rogers", June 3rd, 2014. The invited speakers at the final panel discussion: (from left to right) Bart Lannoo, iMinds, Belgium; Marco Forzati, Acreo, Sweden; Clelia Lorenza Ghibaudo, Telecom Italia, Italy; Thomas Pfeiffer, Alcatel-Lucent Deutschland AG, Germany; Gabriella Cincotti, Università degli Studi Roma TRE, Italy; Domenico Siracusa, Create-NET, Italy; Luca Valcarenghi, Scuola Superiore Sant'Anna, Italy.

NOC 2014 attendees during the visit to models based on da Vinci's engineering sketches at the National Science and Technology Museum Leonardo da Vinci in Milan, June 5th, 2014.

NOC 2014 has been a single track event. The rich technical program of the conference consisted of eight technical sessions and comprised 27 excellent technical regular paper presentations, six invited paper presentations, a keynote speech, and an industrial symposium involving leading experts. Prof. Jaafar Elmirghani (University of Leeds, UK) gave the keynote speech, presening new approaches to minimize power usage in content distribution optical networks. The symposium (organized and chaired by Prof. Pierpaolo Boffi, Politecnico di Milano) focused on flexigrid optical networks, and it was opened by two keynote speeches delivered by Dr. Giuseppe Ferraris (Telecom Italia, Italy) and Dr. Ken Falta (Finisar, USA). Other panelists from Alcatel Lucent, Cisco, and Ericsson joined in the final panel discussion.

The Technical Program Committee (TPC), co-chaired by Prof. Guido Maier (Politecnico di Milano) and Prof. Eiji Oki (University of Electro-Communications, Japan), was built to guarantee a balanced industrial and research (academic) list of members with a solid background in the conference areas. With this objective, the TPC included a list of 37 outstanding experts from around the world. The acceptance ratio of papers at the end of the reviewing process was 50%. Accepted and presented papers will be published in IEEE Xplore as well as other Abstracting and Indexing (A&I) databases. The Publication Chair was Prof. Giacomo Verticale, Politecnico di Milano.

The technical contribution to NOC 2014 has been quite international. The most represented countries in the papers' authorship were: Italy (25%), Spain (11%), France (9%), and Japan (6%), but there were also authors from the USA, Canada, China, Korea, South America, Africa, and the Middle East. A total of 30 countries were represented by authors who submitted papers.

NOC 2014 has been technically co-sponsored by the IEEE Communications Society and by the IEEE Italy Section Photonics Society Chapter, AICT – Association for "Tecnologia dell'Informazione e delle Comunicazioni" of AEIT (Associazione Italiana di

(Continued on Newsletter page 5)

REGIONAL REPORT

The Modernization of the Spanish University System: the Campus of International Excellence Program

By Pilar Manzanares-López, Josemaria Malgosa-Sanahuja, Spain

With the goal of developing a national strategy for modernizing the Spanish university system, and in this way to face the challenges set by the "European Modernization Agenda for Universities: Education, Research and Innovation," in 2009 the Government of Spain started an action plan known as "2015 University Strategy." Presumably, the objectives of this action plan will be evaluated next year to measure the degree of its implementation.

This strategy looks for a significant improvement in the competitiveness of the teaching and research staff as well as to improve the international visibility and academic leadership of Spanish universities. In fact, it follows the model initiated some years before in France, Germany, and the United Kingdom. In relation to university management, this strategy aims to develop an efficient and effective model, accountable to external institutions about the development of its functions. A cornerstone of this strategy is the Campus of International Excellence program (from the Spanish Campus de Excelencia Internacional (CEI)), whose main objective is to improve the internationalization and specialization of Spanish university campuses. This project promotes the evolution of university campuses with the objective of positioning the CEIs among the most prestigious international campuses, and as international references in each of their areas of specialization. This initiative aims to help the Spanish university system improve the quality of its offering and promote efficient and effective teaching and research by means of the promotion of strategic partnerships with institutions, research centers, and companies.

The CEI project has set the targets of enhancing the quality and excellence in teaching and adapting the university studies to the requirements of the European Higher Education Area (EHEA), with particular attention to internationalization and excellence in the field of postgraduate education consisting of masters and Ph.D. programs of international excellence. The CEIs include actions aimed at the recruitment and retaining of international postgraduate and postdoctoral talent, the increase of university education taught entirely in English, the internationalization of the students by means of university exchange programs such as the Erasmus program and other international scholarship programs, the development of training programs oriented to entrepreneurship, the integration of this skill to the curriculum, and the strengthening of the Spanish language courses for foreigners.

In addition, since Spanish universities are institutions where a large part of the Spanish R&I activities are developed, another key objective is to consolidate these institutions as research centers of international excellence and reference from which the obtained knowledge is transferred to society and the productive and business sector.

To reach these goals, the CEI program has considered cutting-edge research, the research activities carried out on the basis of public-private partnerships by means of strategic alliances with research institutes, agencies, and centers of international excellence in the corresponding areas. Therefore, a Campus of International Excellence usually will be composed of the aggregation of more than one university (and each university can be composed of different campuses), public research entities, technology centers, science and technology parks, hospitals, and companies. These strategic alliances seek to provide the transfer of the knowledge and technology resulting from the academic research to the business sector, substantially strengthening the university-society-business relationship.

Another strategic objective of the CEI program is campus transformation, changing the traditional vision of the university campus as a group of buildings with equipment that only offer internal functions to the university community, to a wider vision, inclusive and integrated with the urban environment. The actions proposed in this field include the use of spaces (buildings, student halls of residence, auditoriums, etc.) for cultural activities and public events open to the city, the improvement of sport facilities, and the promotion of community participation in sport events and competitions. Sustainability is another line of action in the projects, promoting actions related to accessibility and mobility, and oriented to the achievement of a sustainable end eco-efficient campus.

The first public call of the CEI program awarded the Campus of International Excellence stamp to nine proposals. A second public call of the same program awarded 14 new proposals. Currently, almost the entire Spanish university system is framed within a Campus of International Excellence.

In particular, the University of Murcia (UM) and the Technical University of Cartagena (UPCT) lead the campus Mare Nostrum 37/38 (CMN). CMN was developed with the vision to become an international resource in the Mediterranean basin to serve as a catalyst for higher educational excellence and a promoter of sustainable product development, using its territory as a model of regional cohesion, and promoting modernization and innovation in key productive sectors of the region that are characteristic of the Mediterranean. The CMN has already established agreements with several Mediterranean Universities and Research Centers of Cyprus, Croatia, Slovenia, Greece, France, Italy, Turkey, Morocco, and Tunisia. The efforts of the CMN are focused on three main research and academic lines: health technologies and their implications in the Mediterranean society's quality of life; naval and marine technologies in the Mediterranean Sea; and the proposal of an eco-economy based on the agro-industry (a filed in expansion in all Mediterranean countries but especially in the north of Africa).

FUSECO/*Continued from page 2*

varying depending on the different starting points and specific targets, making it difficult to compare the different existing initiatives. Essential for any success of Smart Cities are political drive, regulatory and legal support in regard to security and privacy aspects, as well as economic operation and sustainability. All this is demanding for "standardized" and open solutions, which demand for Smart City reference frameworks and architectures and common reusable building blocks. Smart Cities are fueled by a major convergence of quite different application domains and heterogeneous ecosystems into a common framework. This convergence goes far beyond the convergence we have witnessed in the last decades, with many challenges and many opportunities.

Outlook

In the end, Prof. Magedanz and Mr. Hernady closed the Forum by summarizing the key findings and thanking all speakers, chairmen, attendees, sponsors, and local organizing committee for making this forum possible and a success. Most important was the final announcement, that based on this successful first FFAsia, the second FUSECO Forum Asia should be held again in Bali in mid 2015, which by then will also feature an additional IEEE FUSECO Workshop. For more information about FFAsia go to www.fuseco-forum.asia. However, for those interested in the subject matter, the parent event, FOKUS FUSECO Forum, will take place in Berlin, Germany on November 13-14, 2014. See www.fuseco-forum.org for details.

MEMBERSHIP PROGRAMS/*Continued from page 1*

and recognizing the 50th anniversary of ASCII coding as the theme for the 2013 Section banquet. Special guests from industry (the President of Bell Labs was the keynote speaker) and IEEE (past IEEE and ComSoc presidents from the local area) participated as part of this celebration. It was an informative and engaging evening with over 110 attendees.

•The full-day Advanced Communications Symposium was held on Sept 21, 2013 at Stevens Institute of Technology. The ComSoc chapter financially co-sponsored the symposium and chapter members served on the organizing committee.

Bregni: May Chapters receive funding for running their meetings and events?

Hartmann: IEEE MGA provides a small amount of funding to each Chapter yearly, based on a set of minimum performance requirements. In addition, ComSoc may provide funding based on the past year's activity performance and specific programming plans for the coming year. These funds are limited and hence are dedicated to those Chapters that are providing worthwhile technical opportunities to their members. Most chapters do obtain additional support from local companies or universities, such as free meeting space or financial contributions.

Bregni: Merrily, it's almost one year already that you have been serving as NA Director. From your privileged perspective, what are the biggest challenges that ComSoc Chapters have to face in North America?

Hartmann: Major challenges lie in creating a vibrant chapter community with high member retention, membership growth, and individuals willing to take on chapter leadership roles. Online resources (including IEEE Xplore®) make it easier and easier for our members (and potential members) to obtain the information they need without leaving their office/home. Moreover, companies and universities purchase corporate access to IEEE Xplore®, which eliminates part of the value of personal IEEE membership. Job and family time constraints reduce the amount of time for local chapter activities. Many Chapter Chairs have held their positions for several years, due to the fact that no one else is willing or able to step in.

Bregni: Similar challenges are faced by Chapters in all Regions. How do you propose to address them? What would be your recipe?

Hartmann: We try to engage all Chapters and encourage activity. The North America Region Board is made up of representatives from each Region 1-7. In addition, there is a Board member dedicated to coordinating our Distinguished Lecturer/Speaker Tours. We ensure that the benefits of ComSoc membership are made known to our Chapters to assist them in organizing activities. Addressing the specific benefits of ComSoc membership is the mission of Koichi Asatani, Director of Membership Programs Development, and his Board. They provide access to Chapter funding, Chapter Awards, Membership Development Support Grants, the Distinguished Lecturer/Speaker Program, and Student Travel Grants. Moreover, Ashutosh Dutta, Director of Marketing & Industry Relations, and his Board are leading an initiative to assist our Chapters in drawing more members from industry into their activities.

Bregni: In conclusion, what are your plans for 2015?

Hartmann: I would like to focus more on social media for managing our chapters to stay in step with the way our members conduct their daily lives. Our ComSoc IT staff stands at the ready to assist chapters in this regard. Also, our DLT Program delivers timely technical presentations for our Chapters, which help our ComSoc members keep abreast of research and developments impacting their lives and careers. Unfortunately, our DLT budget is exceeded by the annual demand. Therefore, we plan to ask our Distinguished Lecturers to consider also virtual tours (webinars). The ultimate goal, of course, is to make Chapter activities attractive, meaningful, and useful to IEEE and ComSoc members. This is key to increasing ComSoc Chapter membership and to making the Chapter communities active technical environments.

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

JOSEMARIA MALGOSA SANAHUJA, SPAIN (JOSEM.MALGOSA@UPCT.ES)
EWELL TAN, SINGAPORE (EWELL.TAN@IEEE.ORG)

A publication of the IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

NOC 2014/*Continued from page 3*

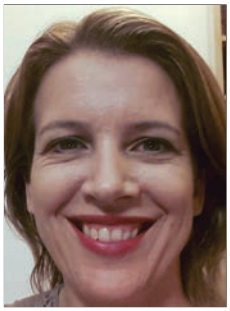
Elettrotecnica, Elettronica, Automazione, Informatica e Telecomunicazioni). The conference had Finisar as a financial patron, and received the patronage of EXPO 2015 (by means of the committee "Le Università per EXPO 2015"). Fondazione Politecnico provided support for the financial management.

It is worth mentioning that NOC 2014 was co-located with three workshops, held on June 3rd on the Politecnico campus "Leonardo," namely: "Optical Access Network Virtualization: from Unbundling to Open Access" (co-organized with the Italian national research project ROAD-NGN), "New Telecom Network Architectures for the Cloud Era" (co-organized by several European FP7 projects), and "Innovation Protection in the ICT" (co-organized by NOC 2014 and AICT).

The leitmotiv for the social events has been the multi-faceted inventor Leonardo da Vinci, who spent many years of his life in Milan, where he realized much of his endless genius production. Before the welcome reception, NOC 2014 participants were guided to visit Politecnico historical Campus "Leonardo" located in Piazza Leonardo da Vinci. The venue chosen for the banquet was the National Science and Technology Museum Leonardo da Vinci, and a visit to the museum before dinner included several models of machines based on da Vinci's engineering sketches. A post-conference guided tour to Biblioteca Ambrosiana was organized on the final day, in which visitors could enjoy a look at the original pages of the Atlantic Codex, one of the most important sets of drawings by Leonardo. NOC 2014 visitors could also admire several masterpieces of Italian Renaissance painting hosted in the adjacent Pinacoteca Ambrosiana.

Finally, we wish to acknowledge the many other persons involved in the event who made it successful thanks to their tireless work: the other chairs, Stefano Bregni (IEEE ComSoc liason) and Massimo Tornatore (workshop organization); the steering-committee members David Faulkner and Alan Harmer; the DEIB personnel; and the local team of postdoc, Ph.D., and master students who assisted speakers and participants.

USER-CENTRIC NETWORKING AND SERVICES: PART 2


Rute Sofia


Alessandro Bogliolo


Fikret Sivrikaya


Huiling Zhu


Olivier Marce


David Valerdi

User-centric networks (UCNs) can be seen as a recent architectural trend of self-organizing autonomic networks where the Internet end user cooperates by sharing network services and resources. UCNs are characterized by spontaneous and grassroots deployments of wireless architectures, where users on such environments roam frequently and are also owners of networking equipment. Common to UCNs is a social behavior that heavily impacts network operation from an end-to-end perspective.

The second part of this Feature Topic starts with an article by T. Jamal *et al.*, "Cooperative Relaying in Dynamic Wireless Networks under Interference Conditions," which describes RelaySpot, a novel link layer relaying protocol based on opportunistic relay selection and cooperative relay scheduling, which shows significant average throughput gains in comparison to proactive opportunistic and broadcast-based relaying approaches.

A second article by Loscri *et al.*, "Spontaneous Smartphone Networks as a User-Centric Solution for the Future Internet," proposes a framework to assist in the deployment of UCNs based on smartphones, which considers, among other aspects, automatic methods to allow smartphones to organize in terms of spontaneous connectivity.

The third article by Mezghani *et al.*, "Content Dissemination in Vehicular Social Networks: Taxonomy and User Satisfaction," provides another perspective on the applicability of UCNs: vehicular social networks. The article contributes a taxonomy for content dissemination as well as proposing the application of utility functions that integrate the notion of user satisfaction as a measure of successful dissemination, that is, the satisfaction level attained by a set of users who receive a specific object.

A fourth article, "A Trajectory-Based Recruitment Strategy of Social Sensors for Participatory Sensing" by Yang *et al.*, represents work developed in the context of relevant services in UCNs — participatory sensing. The article presents a framework for participatory sensing having in mind the goal of developing a trajectory-based recruitment strategy to select social sensors that are well- suited in terms of trust, device temporal availability, and energy consumption.

The fifth and final article of this Feature Topic, "Security and Performance Challenges for User-Centric Wireless Networking" by A. Frangoudis and G. C. Polyzos, identifies challenges of user-centricity in terms of impact on wireless networking architectures, particularly concerning security.

BIOGRAPHIES

RUTE SOFIA (rute.sofia@ulusofona.pt) is director of COPELABS and an associate professor at University Lusofona, Lisbon. She holds a B.Eng. in informatics engineering from the University of Coimbra (1995); and M.Sc.(1999) and Ph.D. (2004) in informatics from the University of Lisboa. Since 1995 she has been developing research in both industry and academia, in packet-based networking, fixed-mobile convergence, and advanced routing/forwarding paradigms. Her research interests comprise pervasive sensing, mobility modeling, and management. She has over 40 publications and 10 patents.

ALESSANDRO BOGLIOLO (alessandro.bogliolo@uniurb.it) is an associate professor of computer systems and coordinator of the School of Information Science and Technology, University of Urbino, Italy. He got a Laurea degree in electrical engineering (1992) and a Ph.D. in electrical engineering and computer science (1998) from the University of Bologna, Italy. He has co-authored over 150 research papers in the fields of low-power electronic systems, wireless sensor networks, bioinformatics, and energy-aware networking.

HUILING ZHU (h.zhu@kent.ac.uk) received her B.S degree from Xidian Univeristy, Xi'an, China, in 1997, and her Ph.D. degree from Tsinghua University, Beijing, China, in 2003. She is currently a lecturer (assistant professor) in the School of Engineering and Digital Arts, University of Kent, Canterbury, United Kingdom. Her research interests are in the area of broadband wireless mobile communications, covering topics such as radio resource allocation and management, MIMO, OFDMA, and cooperative communications.

FIKRET SIVRIKAYA (fikret.sivrikaya@dai-labor.de) is a senior researcher and head of the Network and Mobility group at the Distributed Artificial Intelligence Laboratory (DAI-Labor) of the Technical University of Berlin, Germany. He received his Ph.D. degree in computer science from Rensselaer Polytechnic Institute, New York, in 2007. His research interests include wireless communications, medium access control and routing issues in ad hoc networks, distributed algorithms, and optimization. He has co-authored over 30 peer-reviewed research articles and 6 book chapters.

OLIVIER MARCÉ (Olivier.Marce@alcatel-lucent.fr) joined the former Alcatel R&I in 1999 after two years at the French Research Institute INRIA. Since then, he has been working on packet-based networking, wireless networking, and active networking. His main subjects of interest are related to inter-domain in both wired and wireless, as well as user-defined networks. He has co-authored more than 30 international patents and is the manager of international research projects focused on the integration of IP technologies and wireless networks.

DAVID VALERDI (david.valerdi@fon.com) is responsible for Fon R&D Unit general management. This unit is an active contributor to the Fon technology roadmap by acting as a research center for the latest technology innovations. He was previously technical product manager at Vodafone Group, and also worked for companies like Motorola and Telefónica. He holds several patents and academic publications. He gained telecoms engineering qualifications, at the University of Cantabria, Spain, and received his M.B.A. (with honours) from Instituto de Empresa.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE

INTERNET OF THINGS/M2M FROM RESEARCH TO STANDARDS: THE NEXT STEPS

BACKGROUND

The Internet of Things (IoT) is a framework in which all things have a representation and a presence in the Internet. More specifically, the Internet of Things is aimed at offering new applications and services bridging the physical and virtual worlds, in which machine-to-machine (M2M) communications represents the baseline communication that enables the interactions between Things and applications in the cloud.

The Internet of Things is a key enabler for the realization of the new M2M initiative as it allows for the pervasive interaction with/between smart things leading to an effective integration of information into the digital world. These smart (mobile) things — which are instrumented with sensing, actuation, and interaction capabilities — have the means to exchange information and influence real-world entities and other actors of such a digital world ecosystem in real time, forming a smart pervasive computing environment. The objective is to achieve global access to the services and information through the so-called Web of Things, as well as efficient support for global communications.

The first generation of IoT/M2M standards (from IEEE, IETF, 3GPP, IETF, oneM2M, etc.) is sufficiently mature to enable large-scale operational deployments. Connectivity, vertical data models, RESTful APIs, and device life cycle management are concrete examples of these foundation standards representing the basis of current deployments such as eHealth, automotive, advanced metering infrastructure (AMI), and smart cities.

As the IoT deployment pace accelerates, a next generation of standards is needed to realize the full IoT/M2M vision. This Feature Topic seeks both mature and early research on candidate standards that will transfer to standards organizations such as oneM2M, IETF, 3GPP, IEEE, HGI, and BBF.

Submitted papers in this Feature Topic are expected to focus on state-of-the-art research in various aspects of IoT/M2M from academics and industry viewpoints. The aim of this Feature Topic is thus to offer a venue where researchers from both academia and industry can publish premier articles on the recent advances in theory, application, and implementation of IoT/M2M standards concepts. Topics of interests include, but are not limited to, the following areas of standards research:

- Lightweight protocols and structured data such as Efficient XML Interchange (EXI) and JavaScript Object Notation (JSON) for the IoT
- Interworking with other technologies and systems such as network functions virtualization (NFV) and cloud computing
- Advanced indexing, naming, and addressing of the IoT
- Optimization and enhancement of the currently standardized IoT architectures
- Novel concepts for sensors and actuators such as crowd sourcing
- Abstraction and semantics technologies for devices and services
- Experiences and field trials of IoT applications: smart cities, digital
- IoT/M2M management: Device management evolutions, autonomous management, conflict management, service harmonization
- Next generation of open platforms and hardware for the IoT
- Security, trust, privacy, and iIdentity in the IoT

SUBMISSION GUIDELINES

Articles should be tutorial in nature and written in a style comprehensible to readers outside the specialty of the article. References should be limited to 10, figures and tables to a combined total of 6, and mathematical equations should be avoided. Paper length should not exceed 4500 words. Complete guidelines can be found at http://www.comsoc.org/commag/paper-submission-guidelines. All articles must be submitted through the IEEE Manuscript Central web site (http://mc.manuscriptcentral.com/commag-ieee) by the submission deadline. Submit articles to the category "August 2015/Internet of Things/M2M from Research to Standards."

SCHEDULE FOR SUBMISSIONS

Manuscript Submission: January 15, 2015
Notification of Acceptance: April 1, 2015
Final Manuscript Due: June 1, 2015
Publication Date: August 2015

GUEST EDITORS

Omar Elloumi
Alcatel-Lucent, France
omar.elloumi@alcatel-lucent.com

JaeSeung Song
Sejong Univ., South Korea
mailto:jssong@sejong.ac.kr

Yacine Ghamri-Doudane
Univ. de la Rochelle, France
yacine.ghamri@univ-lr.fr

Victor Leung
University of British Columbia
vleung@ece.ubc.ca

Cooperative Relaying in User-Centric Networking under Interference Conditions

Tauseef Jamal and Paulo Mendes

ABSTRACT

An ever-growing demand for pervasive Internet access has boosted the deployment of wireless local networks in recent decades. Nevertheless, wireless technologies face performance limitations due to unstable propagation conditions and mobility of devices. In face of multi-path propagation and low-data-rate stations, cooperative relaying promises gains in performance and reliability. However, cooperation procedures are unstable, due to their dependence on current channel conditions, and introduce overhead that can endanger performance, especially when nodes are mobile. This article presents an introduction to cooperative relaying, and describes a novel link layer protocol, called RelaySpot, able to implement cooperative relaying in dynamic networks, based on opportunistic relay selection, cooperative relay scheduling, and switching.

INTRODUCTION

The growth of wireless networks in the last decades is motivated by their ability to support communications anywhere and any time. Boosted by the importance that such pervasive communications have on modern society, a high proliferation of wireless services and devices, such as WiFi and wireless phones, has emerged. The increasing deployment of wireless systems, in particular by end users (e.g., wireless home gateways, smart phones, wireless embedded devices) brings new challenges to the deployment of reliable wireless systems due to variations in the density of wireless networks, the unpredictable coverage of such networks, and the unstable availability of wireless devices (e.g., due to users' mobility and patterns of use).

User-centric networking is a new trend that aims to support efficient communications in such dynamic wireless environments by exploiting the role that end users may have in the networking process by sharing network services and resources. This user-centric perspective of networking opens new possibilities for the development of novel wireless technologies able to sustain reliable and cost-efficient wireless transmissions, even when wireless nodes are pervasive and mobile, and wireless communications are subjected to correlated interference conditions. One such wireless technology is cooperative relaying.

Cooperative relaying aims to bring several improvements to wireless networks, from increasing network capacity and coverage, to enhancing transmission reliability and throughput even in scenarios where mobile devices communicate under different wireless interference conditions [1].

In this article, we start by providing an introduction of the role that cooperative relaying may have in the development of efficient user-centric wireless networks. In such networks cooperation occurs when overhearing wireless devices (called relays) in different locations assist the communication between source to destination by transmitting different copies of the same frame, generating spatial diversity that allows the destination to get independently faded versions of the transmitted message. The literature reveals different cooperative approaches, depending on the role that sources, relays, and destinations have on the cooperation process. In this article we argue that the most suitable approach is the one that presents more self-organized properties in order to be efficient in dynamic wireless scenarios. For this purpose, we present RelaySpot as a solution to mitigate the problems posed by wireless fading and low-data-rate mobile devices. With RelaySpot, relays are self-elected if within a cooperation area, defined for a source-destination pair, they overhear a good frame transmitted by the source. Destination nodes are able to select the best set of relays based on the information provided by the latter during a predefined time period.

This article is organized as follows. We provide an introduction to cooperative relaying in general. We provide an analysis of cooperative relaying approaches, and we describe the proposed RelaySpot protocol. We present experimental analysis. Finally, we present our conclusions.

The authors are with COPELABS/University Lusofona.

0163-6804/14/$25.00 © 2014 IEEE

Cooperative Relaying: Advantages and Limitations

Over the past decade, Internet access has become essentially wireless, with 802.11 technologies providing low-cost broadband support for flexible and easy deployment. However, channel conditions in wireless networks are subject to interference and fading, decreasing the overall network performance [2]. While fast fading can be mitigated by having the source retransmit frames, slow fading, caused by obstruction of the main signal path, makes retransmission useless, since periods of low signal power last for the entire duration of the transmission. Moreover, interference from other transmitters also affects the communication quality. Due to continuous changes related to interference conditions and mobility of devices (transmitter, receiver, or relay), the wireless signal is scattered over many surrounding objects. Such channel impairments can be mitigated by exploiting cooperative diversity [3].

Besides the limitations inherent in the wireless channel, wireless networks suffer from, among other issues, scarcity of bandwidth, which limits the network throughput and requires efficient utilization of this resource. One example is the system impairment caused by the presence of low-data-rate devices. The 802.11 standard makes use of rate adaptation schemes to allow low-data-rate devices to adapt their modulation and coding scheme according to the quality of the radio channel, improving the bit rate and robustness of their data transmission. However, the usage of rate adaptation schemes results in degradation of the overall network performance, since low-data-rate devices grab the wireless medium for a long time. This occurs when each device has the same probability to access the wireless channel, which means that high-data-rate devices are not able to keep the desirable throughput. Cooperative relaying may mitigate this problem by allowing low-data-rate devices to finish their transmission faster by using a pair of wireless links (via a relay) that provide better wireless conditions than the direct channel to the destination. High-data-rate devices have a high incentive to cooperate by relaying messages from low-data-rate devices, since such cooperation may increase their probability to grab the wireless channel faster.

For a better understanding of the role of cooperation in wireless networks, Fig. 1 shows a basic 802.11b system where devices have different transmission rates at different distance from the access point (AP). Although cooperative relaying may improve the performance of such a heterogeneous system, such improvement can only be ensured if the relay device is within a cooperation area that rectifies the impact of low-rate nodes. Figure 1 provides an example where a source, placed at a distance from the AP that only allows it to transmit at 2 Mb/s, can actually transmit at 11 Mb/s, by making use of any relay placed in a suitable cooperation area. Such a cooperation area is identified as the interception of R_{11}, referred to the distance at which the AP can receive at 11 Mb/s, and r_{11}, which is the distance at which the source can transmit at 11 Mb/s.

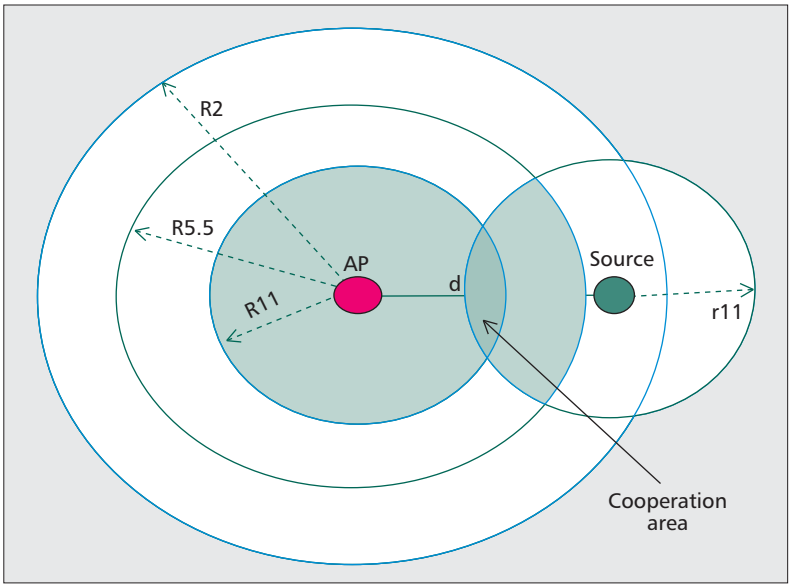


Figure 1. Cooperation conditions in wireless networks.

Although cooperative relaying brings advantages for wireless networks, it is necessary to analyze the potential drawbacks: one is related to the additional interference caused by the relaying operation, since it involves additional transmissions via relays. Thus, the benefits brought by cooperation can be diminished if the relaying mechanism is not cleverly designed. Other potential constraints are concurrent transmissions and mobility, which can affect the performance of cooperative networks [1]. Therefore, there are design issues that must be taken into account while developing cooperative systems aiming to exploit wireless diversity at the link layer, such as relay selection, cooperation decision, cooperation notification, cooperative transmission, and cooperation management [1].

Of all the design issues, relay discovery and selection is of high importance. Concerning relay selection, most of the existing protocols require some devices, normally the source, to have a neighborhood map related to channel conditions. Such a map is normally updated based on a broadcast mechanism: broadcasts need to be very frequent to cope with network variations, which limits the performance of the wireless system.

Besides the decision about the best relay, or set of relays, to use, it is necessary to keep cooperation efficiency. This aspect is of high importance in dynamic scenarios where mobile devices face variant interference conditions. Concerning relay management, most of the protocols use additional control messages in a centralized manner. Such explicit notifications affect the cooperation gain due to extra overhead. Moreover, in some scenarios, it is infeasible to have centralized coordination [4]. The challenge here is the development of a distributed relay switching mechanism that allows the wireless system to take advantage of the most suitable relaying conditions.

As described in this section, the limitations of

One advantage of opportunistic relaying is its independence from any neighbor map maintained by means of extra exchanges of messages. This property allows a relay to forward data opportunistically without prior coordination among a set of devices.

the cooperation process can be as significant as its advantages. Therefore, cooperative network design needs to be performed carefully in order to achieve the full gains of cooperation while ensuring that cooperation does not cause degradation of system performance. The next section provides an analysis of different types of proposals aimed at identifying the most suitable approach for relaying in dynamic networks facing variant interference conditions.

Analysis of Cooperative Relaying Approaches

In general terms, cooperative relaying at the link layer comprises two phases: relay selection and cooperative transmission. In the first phase a relay or group of relays are selected, while in the latter phase the transmission via relay(s) takes place. The relays can be selected by either source (source-based), destination (destination-based), or the relay itself (relay-based). At the link layer we can classify cooperation protocols as proactive and reactive. With proactive relaying, cooperation is set up to improve the performance of the direct link, even if the latter is operational [5]. In proactive relaying the cooperation process can be controlled by the source, destination, or potential relay. Proactive relaying is time-critical and incurs in higher overheads: frequent information exchange for timely delivery of data is required. In reactive relaying the cooperation is initiated when the direct link is not operational, which can be detected by any device receiving or overhearing a negative acknowledge message, or a lack of communication, which can occur due to collision or transmission errors [1].

Reactive relaying incurs in lower overhead, but is only appropriate for applications that are tolerant to delays or disruption. As mentioned, cooperative relaying approaches can be classified as proactive and reactive, proactive and reactive relaying can be further divided into broadcast-based, and opportunistic. Broadcast-based approaches represent a relatively simple strategy by making use of the broadcasting nature of the wireless medium. While broadcast-based relaying offers more control due to its centralized nature, with opportunistic relaying devices can make cooperation decisions on their own, within certain time and spatial constraint. As a general property, opportunistic relaying does not require extra control messages. The rest of this section aims to highlight the differentiation factor of applying a broadcast or opportunistic approach to relaying.

Broadcast-Based Relaying

Broadcast-based relaying relies on the existence of a neighborhood map of channel conditions, normally at the source or destination. The major drawback of broadcast-based relaying is the periodic broadcasts required for maintaining the neighborhood map and the consequent extra control overhead, which affects the performance. As mentioned before, broadcast-based approaches can be implemented in a proactive or reactive fashion. These approaches are normally source- or destination-based.

One example of proactive source-based cooperative relaying scheme at the link layer is the Cooperative Medium Access Control (CoopMAC) protocol [6]. With CoopMAC the source selects (source-based) an intermediate device (relay) that has a relatively good channel between the source and the destination. Based on the channel state information (CSI) broadcast by potential relays, sources update a local table (cooptable) used to select the best relay for each transmission. CoopMAC performs a three-way handshake, which requires the selected relay to send a control message, called helper ready to select (HTS), between the request to send (RTS) and clear to send (CTS) messages: first, the source sends a cooperative RTS (CoopRTS) message with the selected relay ID. If the selected relay is willing to cooperate, it then sends an HTS message back to the source. If the destination overhears an HTS message, it transmits a CTS. After receiving a CTS, the source sends the data frame to the destination via selected relay.

Opportunistic Relaying

One advantage of opportunistic relaying is its independence from any neighbor map maintained by means of extra exchanges of messages. This property allows a relay to forward data opportunistically without prior coordination among a set of devices. Hence, such approaches are normally relay-based. Opportunistic relaying is suitable for the deployment of cooperative relaying in dynamic scenarios. However, opportunistic relaying presents some drawbacks, such as relays backing off every time they forward; and the source ignoring the availability of potential relays, hence not knowing the data rates of source-relay and relay-destination channels.

One example of opportunistic (reactive) relaying is cooperative Ccommunication MAC (CMAC) [7]. In CMAC each device stores the data frames sent by source. If no acknowledgment (ACK) is overheard, the relay forwards the stored data frame on behalf of the source. Due to usage of additional queues and channel estimations, CMAC introduces extra overhead.

RelaySpot: A Hybrid Relaying Solution

RelaySpot is a hybrid cooperative relaying protocol where relays are self-elected under certain cooperation conditions. Selected relays are used to increase the performance of active transmissions (proactive behavior) or replace failed transmissions (reactive behavior). RelaySpot comprises three building blocks: opportunistic relay selection, cooperative relay scheduling, and relay switching. In [3] we presented the concept of the RelaySpot framework.

Moreover, relay selection faces several optimization problems, meaning that the best relay may be difficult to find. Hence, for dynamic scenarios, the approach followed by RelaySpot is to make use of the best possible relaying opportunity, and to switch between relays qualified within the cooperation area if necessary.

Figure 2 illustrates the RelaySpot operation

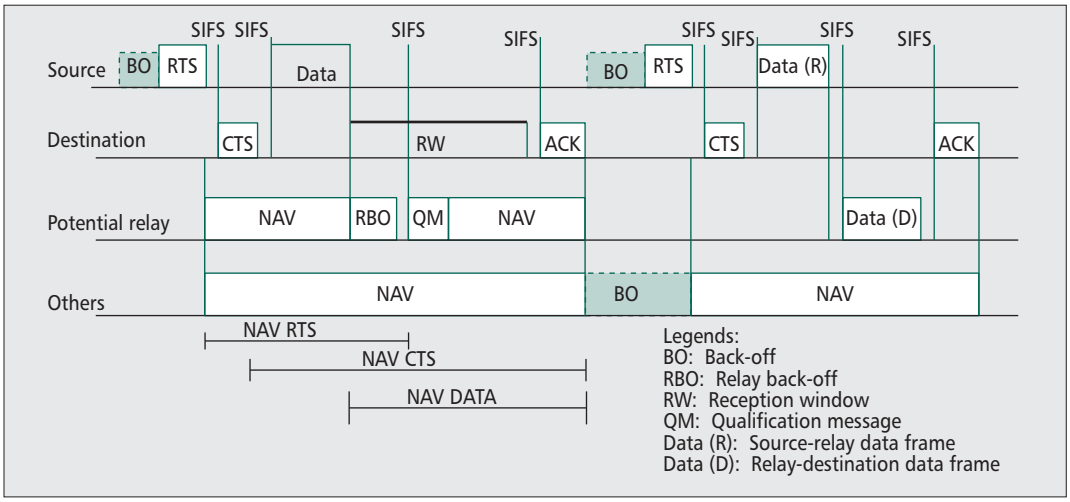


Figure 2. Proactive mode operational example.

in a scenario with a poor direct link between source and destination.

In this scenario, when the destination observes a poor data rate, it implicitly asks for relaying in CTS, since such a message is overheard by any potential relay. As a result potential relays opportunistically start a self-electing procedure. Based on the information sent by self-elected relays to the destination, the latter chooses the best relay or set of relays among self-elected relays (cooperative relay scheduling). The source then starts cooperative transmission, in which it relays the following data frames via that selected relay. This procedure continues until the quality of the direct link improves.

The remainder of this section provides a description of RelaySpot functional components that allow relays to be opportunistically selected and the destination to schedule the potential relays for the forthcoming transmissions. Some of these transmissions may use different relays if a relay presents better conditions than the current one.

Opportunistic Relay Selection

Relay selection is a challenging task, since it greatly affects the performance of a cooperative network. Relay selection may introduce extra overhead and complexity, and may never be able to find the best relay in dynamic scenarios. Hence, the major goal of RelaySpot is to minimize cooperation overhead, with no performance degradation, by defining a relay selection process able to take advantage of the most suitable self-elected relay [8].

With RelaySpot, relay selection is performed in three steps: first, verification of the eligibility of devices to become relays, which occurs if devices are able to overheard a good frame sent by the source and are positioned within the cooperation area; second, computation of the selection factor (SF) of eligible devices; and third, computation of the contention window (CW) of eligible devices based on their SF. At the end of the selfelection process, eligible relays send a qualification message (QM) toward the destination after the expiration of their CW. During the first phase, potential relays (i.e., nodes that overhear both RTS/CTS) are verified if they are inside the cooperation area by computing their cooperation factor (CF), which is related to the effective rate of the source-relay channel (R_{sr}) and relay-destination channel (R_{rd}). These rates are computed by overhearing RTS and CTS frames exchanged between source and destination. The CF ensures that potential relays are closely bounded with the source, while having a good channel toward the destination: an eligible relay must have a CF that ensures a data rate higher than the data rate achieved when using the direct link.

During the second phase, the computation of the SF of a relay depends solely on local information related to interference (node degree and load), mobility, and history of successful transmissions toward the specified destination. Node degree, estimated by overhearing the shared wireless medium, gives an indication about the probability of having successful relay transmissions: having information about the number of neighbors allows the minimization of collision and blockage of resources. However, it is possible that devices with low device degree are overloaded due to:

- Processing demands of local applications (direct interference)
- Concurrent transmissions among neighbor devices (indirect interference)

Hence, RelaySpot relies on node degree and traffic load generated and/or terminated by the potential relay itself to compute the overall interference level to which each potential relay is subjected. The goal is to select as relay a device that has low interference factor, which means few neighbors (ensuring low blockage probability), and fast indirect and direct transmissions (ensuring low delays for data relaying).

By using the interference level together with the history and mobility factors, the probability of selecting a certain device as a relay for a given destination is proportional to the history of successful transmissions the device has toward that destination plus its average pause time, while being inversely proportional to the interference level to which the device is subject.

During the third phase (CW computation),

Relay selection is a challenging task, since it greatly affects the performance of a cooperative network. Relay selection may introduce extra overhead and complexity, and may never be able to find the best relay in dynamic scenarios.

the goal is to increase the probability of having successful transmissions from relays to the destination by giving more priority to relays that are more closely bounded to the destination, and have less interference and higher pause times.

Cooperative Relay Scheduling

The relay selection mechanism may lead to the qualification of more than one device as relay, each with different values of SF, leading to CWs with different dimensions.

The destination schedules the self-elected relays after the expiration of a reception window (RW), in order to receive as many QMs as possible, as illustrated in Fig. 3. The size of the RW is of major importance: on one hand, a large window increases the probability of scheduling a good relay based on a large set of received QMs; on the other hand, a small window introduces a lower delay in the communication session.

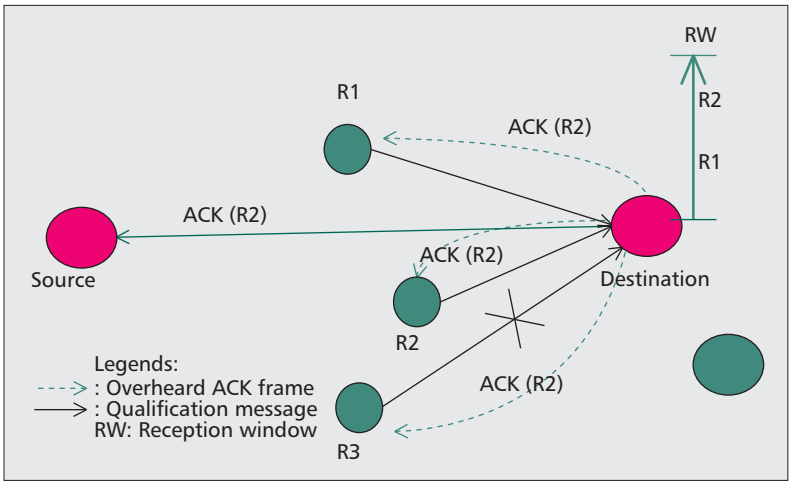


Figure 3. Relay scheduling example.

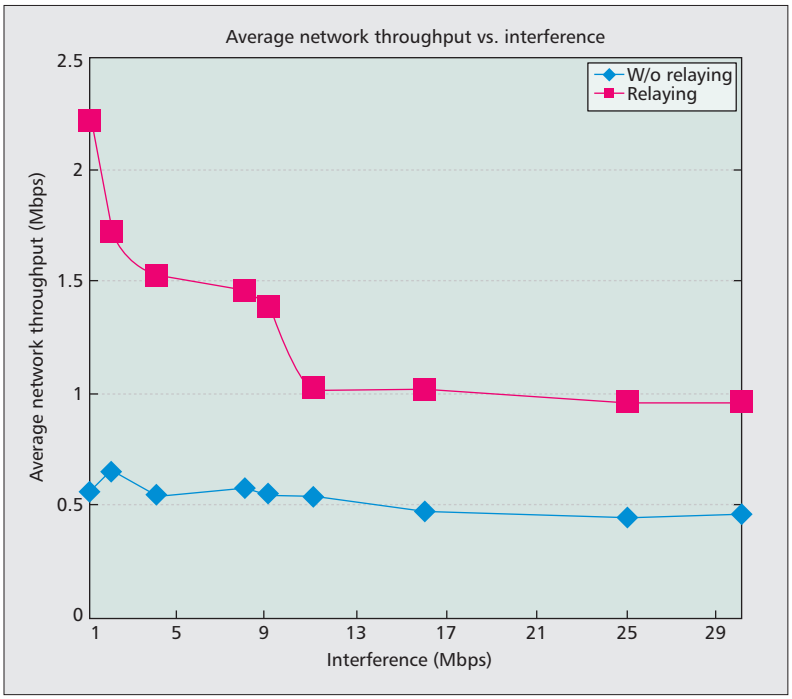


Figure 4. Analysis of the impact of interference.

After the expiration of the RW, the destination schedules all the eligible relays by checking the R_{rd} (by means of the received signal strength) and the R_{sr} (by means of the information carried in the QM).

After the scheduling process, the destination sends an ACK message to the source including the MAC address of the selected relay, which will be responsible for forwarding all forthcoming frames of the communication session to the destination. The ACK frame also piggybacks the CF information.

This cooperative scheduling procedure supports RelaySpot proactive operation, by having the destination only schedule relays that present a combination of R_{sr} and R_{rd} with better data rate than the direct link. In RelaySpot the proactive operation is complemented by a reactive procedure, in which the decision on the cooperative scheduling mechanism is overtaken by having another relay, in the cooperation area, replace the relay previously selected by the destination when the latter fails. This relay switching mechanism is described below.

Relay Switching

Since relays are selected opportunistically based on local information, there is a probability of having good relays computing CWs that are not small enough to allow them to send a QM before the expiration of the RW. In order to overcome this situation, as well as to support the failure of selected relays, RelaySpot includes a relay switching operation able to select one relay among a set of potential relays able to cooperate when needed.

All potential relays are able to compute their own CF, as well as the CF of the selected relay, by overhearing ACK frames.

If a potential relay is not selected in the relay selection procedure, it compares its CF with the CF of the selected relay. If its CF is better, meaning that it can provide better performance gain to the ongoing communication session, it sends a switching message to the destination by means of a dummy data frame, informing it of its own CF. This way, the selected relay can be switched to the newly relay, since:

- By overhearing the frame sent by the new relay, the source will send the next data frame toward that relay.
- By receiving the frame sent by the new relay, the destination knows that the next data frame will be sent by it.

Relay switching is suitable for dynamic scenarios where a previously selected relay may not be efficient at some stage (e.g., due to mobility, fading, or obstacles). Hence, unlike prior art, relay switching can overcome such variations in network conditions, making the deployment of cooperative relaying possible for dynamic networks.

Relay switching can be used to improve the performance of a communication session by replacing a good relayed transmission by a better one, as well as to implement a reactive operation. The latter is implemented by having relays switched implicitly when a potential relay detects a missing ACK frame for an already relayed communication. In this situation, the potential

relays try to forward the overheard data frame on behalf of the relay that failed the transmission. In case of success, the destination notifies the source, by means of an ACK frame, about the MAC address of the potential relay that first reacted to the failure of the relayed communication.

Experimental Analysis

RelaySpot evaluation is based on simulations run on the MiXiM framework of the OMNeT++ 4.1 simulator using a 2D linear mobility model. Simulations consider a scenario with one static AP and up to 25 mobile devices in a area of 200 × 200 m^2. In this section we start by presenting a study of the performance of RelaySpot in scenarios with different levels of interference and mobility. This study is done against standard 802.11 and a mobility-unaware version of RelaySpot. Second, we present a comparative study of RelaySpot against broadcast-based and opportunistic relaying approaches.

Impact of Interference

Figure 4 illustrates the advantages of using RelaySpot as a complement to normal 802.11 operation in scenarios with high interference. In this experiment, we consider a scenario where one static source is placed at enough of a distance from the AP to observe a poor data rate; interference is added by randomly placing transmission pairs (each with 5 Mb/s on average) among the 25 available relay nodes. In this experiment RelaySpot is configured without relay switching, since the object is to analyze the efficiency of RelaySpot in selecting and scheduling a good relay in the presence of interference.

Simulation results show that in the presence of interference, RelaySpot has better performance than IEEE 802.11 (147 percent higher throughput than the standard 802.11 on average), by avoiding selecting overloaded devices as relays, and selecting relays with low blockage probabilities and with good transmission success rate toward the destination [9].

Network Capacity Analysis

In this section we analyze the performance of RelaySpot based on its impact on the overall network capacity (i.e., average network throughput) of a wireless network. In this experiment we consider a scenario with 25 mobile devices moving with random pause time between 10 and 100 s. The goal is to understand if RelaySpot can increase the capacity of the network by increasing the overall average throughput in the presence of devices with different levels of mobility. In this experiment, we compare the average network throughput achieved by a version of RelaySpot without mobility awareness against 802.11.

Simulation results (Fig. 5) show that RelaySpot can achieve higher throughput than 802.11 and mobility-unaware RelaySpot even with high load. The main reason is that RelaySpot is able to select relays with high pause time, which reduces the overall communication delay by avoiding reselection of relays during the communication session.

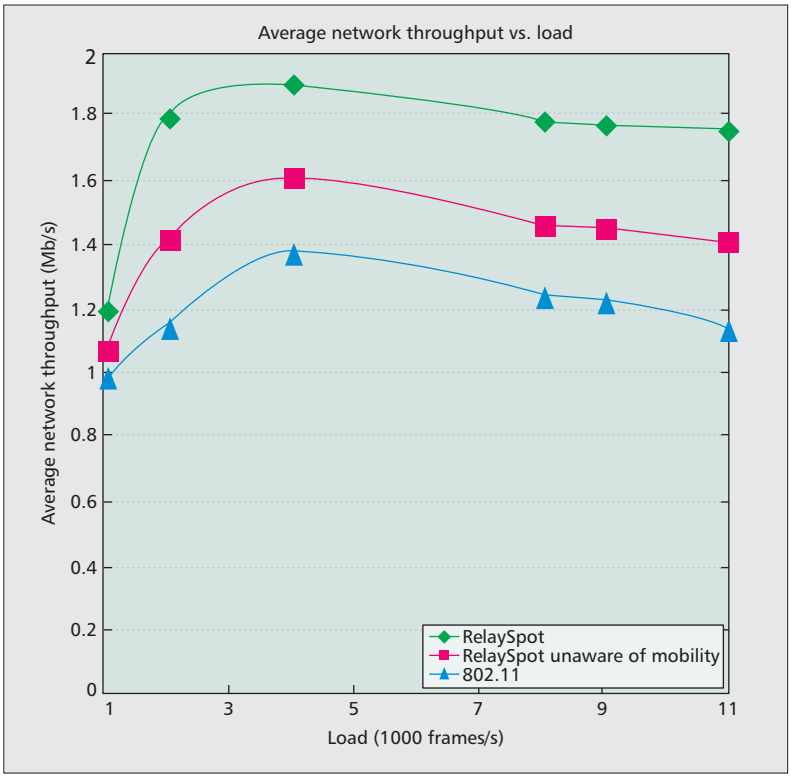


Figure 5. Impact of RelaySpot on overall network capacity.

RelaySpot achieves an average throughput gain of 42 percent in relation to 802.11, and 21 percent in relation to the RelaySpot version that is unaware of mobility. Without mobility awareness, RelaySpot can still achieve an average throughput gain of 17.6 percent in relation to 802.11 due to the scheduler at the destination, which is able to select a relay with a pair of channels (source-relay; relay-destination) with better throughput than the direct link even in the presence of mobile devices.

Comparative Analysis

This section provides an analysis of the hybrid relaying approach, followed by RelaySpot, against two generic implementations of proactive opportunistic and broadcast-based approaches [10]. We perform simulations using network load of 10,000 frames/s.

Figure 6 shows a clear advantage of using a hybrid approach in dynamic networking scenarios due to its capability to react to relay failures by exploring a relay switching functionality. Relay switching is able to decrease the overall contention by avoiding relay reselection and replacing relays with poor performance.

Broadcast-based relaying includes additional control messages for handshake to avoid collisions and guarantee correct channel reservations. This is why it achieves an average throughput gain of 40 percent in relation to 802.11. However, the gain decreases with the increase of network density, since relay failure increases due to collisions.

Both hybrid (RelaySpot) and broadcast-based relaying achieve better throughput gain compared to opportunistic-based relaying. Figure 6

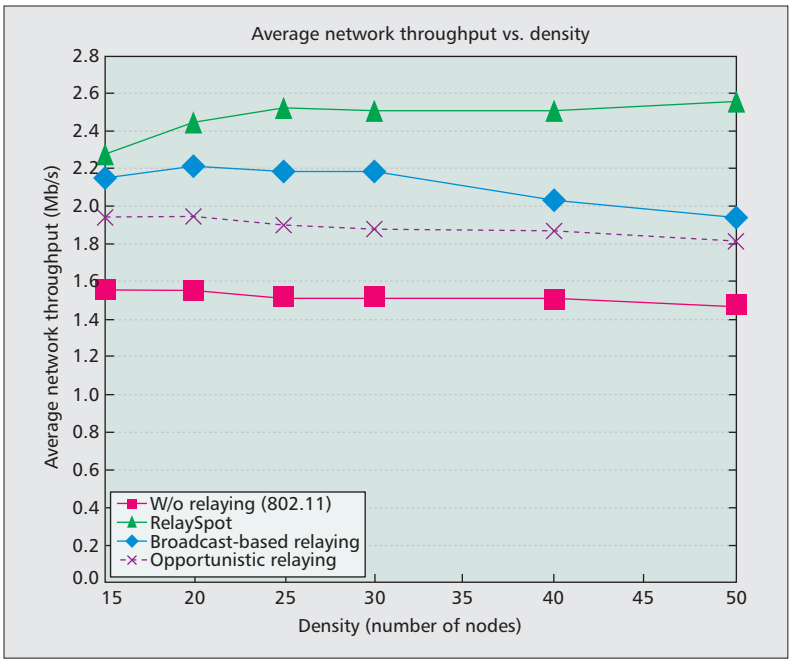


Figure 6. Comparative analysis of hybrid, broadcast, and opportunistic relaying.

shows that opportunistic relaying achieves an average throughput gain of only 24 percent in relation to 802.11. Such low gain is due to the fact that the source and destination do not know the availability of relays, leading to a high probability of a failed relay attempt and collision.

Figure 6 shows that hybrid relaying, such as RelaySpot, are able to increase the overall network performance, while decreasing the impact of relaying overhead. Broadcast-based and opportunistic relaying lead to decreased overall network throughput with increased density. The main reason is its ability to select good relays on a first attempt (e.g., relays with low interference and low mobility), as well as its ability to quickly replace relays that start to present poor performance.

Conclusion

The user-centric perspective of networking opens new possibilities for the development of novel technologies able to sustain reliable and cost-efficient wireless transmissions, such as cooperative relaying. Cooperative relaying aims to bring several improvements to wireless networks, from increasing network capacity and coverage to enhancing reliability and throughput even in scenarios where mobile devices communicate under different wireless interference conditions.

In this article we argue that hybrid cooperative relaying is the most suitable approach for such dynamic scenarios due to its self-organized properties. To justify our argument, we present RelaySpot as a solution to mitigate the problems posed by fading and the presence of low-data-rate mobile nodes. Experimental results show that RelaySpot can effectively increase the capacity of wireless local area networks even in the present of mobile nodes communicating under different interference conditions. The proposed approach achieves an average throughput gain of 32 and 18 percent in relation to proactive opportunistic and broadcast-based relaying, respectively.

References

[1] T. Jamal, P. Mendes, and A. Zúquete, "Wireless Cooperative Relaying Based on Opportunistic Relay Selection," *Int'l. J. Advances in Networks and Services*, vol. 05, no. 2, June 2012, pp. 116–27.
[2] W. Elmenreich *et al.*, "Building Blocks of Cooperative Relaying in Wireless Systems," *Electrical and Computer Engineering*, Springer, vol. 125, no. 10, Oct. 2008, pp. 353–59.
[3] T. Jamal, P. Mendes, and A. Zúquete, "RelaySpot: A Framework for Opportunistic Cooperative Relaying," *Proc. IARIA ACCESS*, Luxembourg, June 2011.
[4] A. Ozgur, O. Leveque, and D. Tse, "Hierarchical Cooperation Achieves Optimal Capacity Scaling in Ad Hoc Networks," *IEEE Trans. Info. Theory*, vol. 53, no. 10, 2007, pp. 3549–72.
[5] H. Shan, W. Zhuang, and Z. Wang, "Distributed Cooperative MAC for Multihop Wireless Networks," *IEEE Commun. Mag.*, vol. 47, no. 2, Feb. 2009, pp. 126–33.
[6] P. Liu *et al.*, "CoopMAC: A Cooperative MAC for Wireless LANs," *IEEE JSAC*, vol. 25, no. 2, Feb. 2007, pp. 340–54.
[7] S. Shankar, C. Chou, and M. Ghosh, "Cooperative Communication MAC (CMAC): A New MAC Protocol for Next Generation Wireless LANs," *Proc. Int'l. Conf. Wireless Networks, Commun. and Mobile Computing*, Maui, HI, 2005.
[8] T. Jamal and P. Mendes, "Relay Selection Approaches for Wireless Cooperative Networks," *Proc. IEEE WiMob*, Niagara Falls, Canada, Oct. 2010.
[9] T. Jamal, P. Mendes, and A. Zúquete, "Opportunistic Relay Selection for Wireless Cooperative Network," *Proc. IEEE IFIP NTMS*, Istanbul, Turkey, May 2012.
[10] T. Jamal and P. Mendes, "Analysis of Hybrid Relaying in Cooperative WLAN," *Proc. IFIP WirelessDays*, Valencia, Spain, Nov. 2013.

Biographies

Tauseef Jamal (tauseef.jamal@ulusofona.pt) graduated in electronics from the Islamia College University Peshawar in 1996, received his M.Sc. in electronics from the University of Peshawar, Pakistan in 1998, his PGD in computer systems in 2000 and M.S in computer engineering from Halmstad University, Sweden, in 2006, and his Ph.D. in telecommunications from the University of Aveiro, Porto and Minho, Portugal. Currently, he is a researcher in the research unit of Informatics Systems and Technologies (SITI) of University Lusofona within CopeLabs. His major research interests are cooperative networking, cooperative routing, and self-organized wireless networks.

Paulo Mendes (paulo.mendes@ulusofona.pt) is a vice-director of COPELABS , SITI coordinator, and an associate professor of University Lusófona,Portugal, where he heads the Ph.D. program in informatics, NEMPS. He received a B.Eng. in informatics engineering from the University of Coimbra in 1993, an M.Sc. in computers and electrotecnical engineering from IST, UTL, Lisboa, Portugal, in 1998, and a Ph.D. in informatics engineering from the University of Coimbra in 2004. His research and teaching interests are in the area of self-organizing networks, cooperative relaying, and complex networks.

Spontaneous Smartphone Networks as a User-Centric Solution for the Future Internet

Gianluca Aloi, Marco Di Felice, Valeria Loscrì, Pasquale Pace, and Giuseppe Ruggeri

ABSTRACT

In this article we focus on a special case of user-centric networks, spontaneous smartphones-based networks, SSNs, where the role of the end-user devices is played by smartphones that are "evolutionary" and more active in supporting communication services. SSNs present key features like spontaneity in the creation of the network and redefinition of the devices' role in order to make them continuously adaptive to both network and user requirements. This work is devoted to identifying the potential advantages of SSNs by also providing a clear definition of the challenges and issues that need to be faced in order to make this emerging paradigm effective and practically deployable.

INTRODUCTION

Nowadays, the spontaneous smartphone network (SSN) is emerging as a potential new communication paradigm, characterized by the fact that the access network has a strong self-organizing nature and is primarily made up of user-owned devices. These latter can also act as routers by actively cooperating to forward data on multihop paths. SSNs can be considered a special case of user-centric networks (UCNs) made up only of users' devices.

This new approach to network deployment undoubtedly offers great advantages in terms of:

- Lower costs required to set up network access
- Reduced or no maintenance at all for network management
- The possibility to set up a network even in scenarios where the infrastructure is barely available (disaster scenarios, rural areas, least developed countries, etc.)

Despite the enormous potential of SSNs, multihop communication between smartphones or similar devices (e.g., phablet, tablet) is still not an affirmed paradigm, and self-organized SSNs are a challenge [1]. Usually, network creation and management requires massive intervention from users who, however, prefer being agnostic about technological issues. This requirement constitutes a unique issue of SSNs compared to traditional self-organizing systems and generic multihop ad hoc networks. Hence, a framework able to limit human intervention, with the aim of making network management *as spontaneous as possible*, is mandatory for the success of the SSN paradigm.

In this work, we present and discuss results of a novel framework named STEM-Net [2] specifically adapted to SSNs, in order to provide two main enabling features for users' devices:

- Network self-configuration
- Evolvability, that is, the capacity of continuously adapting to the needs of both users and networks

Through STEM-Net, a smartphone can switch between different roles. It can produce/receive data, forward the traffic of other terminals through multihop communications, or provide access to global network resources to other terminals; but above all, it can autonomously assume the most suitable role without any user intervention.

The main contributions of the article are:

- To specify advantages and issues of SSNs by identifying strategic scenarios and use cases on which SSNs can be utilized
- To review the enabling technologies of SSNs by also pointing out limitations of existing software frameworks in terms of self-organizing deployments
- To describe how the application of the STEM-Net framework can solve such issues by also presenting a proof-of-concept implementation on real smartphones

USE CASES AND SCENARIOS

In this section, we propose few communication scenarios (depicted in Fig. 1) well suited for SSNs to motivate the need for a new and flexible system architecture for network creation and management.

Pervasive wireless Internet access: SSNs can be used to deploy fully pervasive and infrastructure-free Internet access by naturally extending the coverage of wide areas. This goal could easily be achieved because of the huge amount and density of smartphones located in the majority of daily environments. As a result, a pervasive sce-

Gianluca Aloi and Pasquale Pace are with DIMES — University of Calabria.

Marco Di Felice is with DISI — University of Bologna.

Valeria Loscrì is with INRIA Lille — Nord Europe.

Giuseppe Ruggeri is with DIIES — University "Mediterranea" of Reggio Calabria.

0163-6804/14/$25.00 © 2014 IEEE

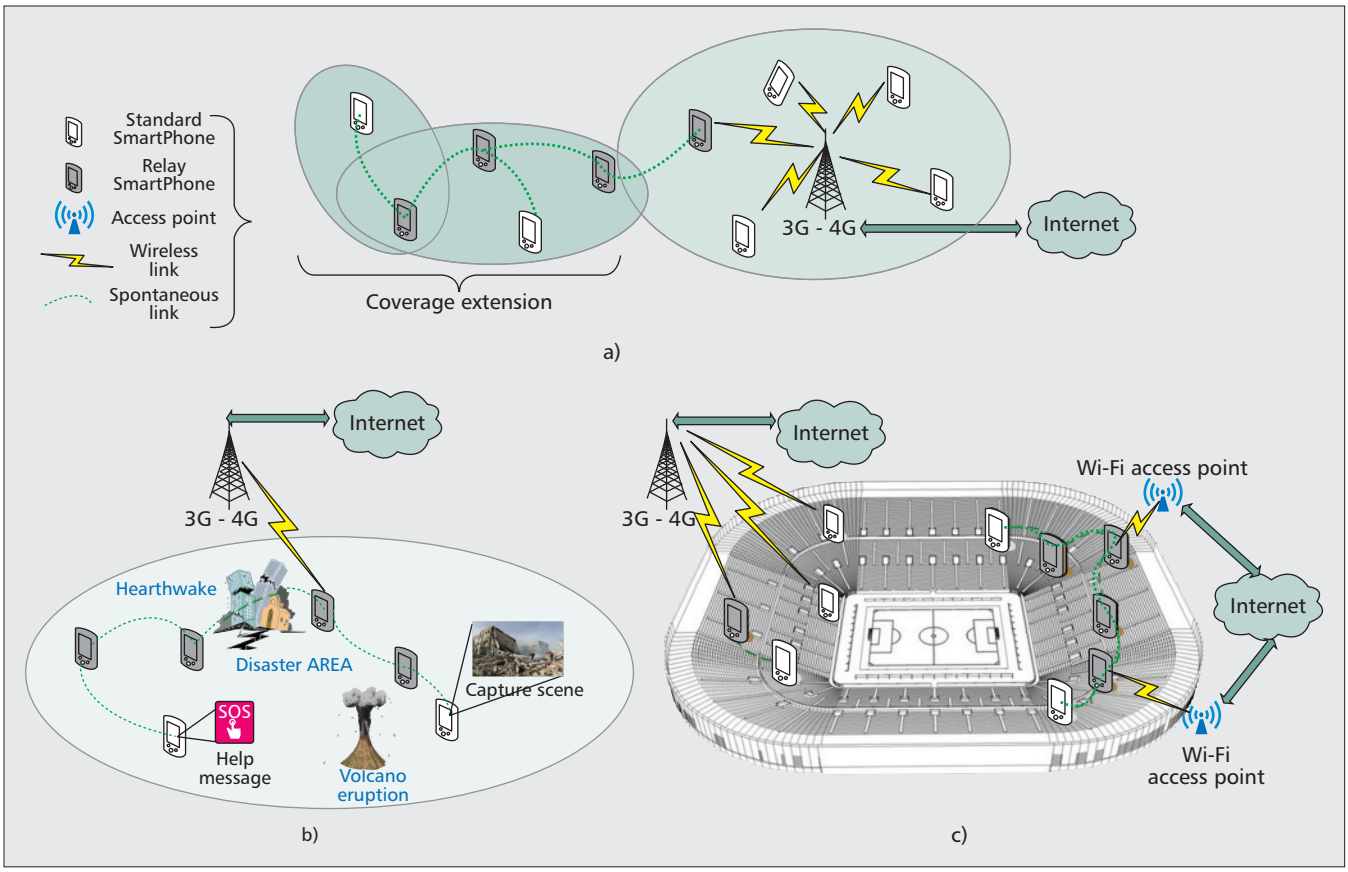


Figure 1. Use cases and communication scenario: a) pervasive wireless Internet access; b) emergency and post-disaster recovery; c) large crowd gathering places.

nario can be implemented by adding self-configuration capabilities to smartphones, making them able to cooperate with the aim of sharing access to global resources.

Emergency and post-disaster recovery: When natural catastrophes disrupt traditional network infrastructures, SSNs can help both survivals and rescue teams. The survivors, by using their mobile phones and the scarce communication resources still available, can share updates, post photos, and upload videos on social media sites. We are confident that in the long run, for emergency services, SSNs will complement the safety networks (prospectively based on Long Term Evolution, LTE) in place only for officials such as police and fire brigades.

Large crowd gathering places: When special events (e.g., concerts, trade fairs, Olympic games) involve huge numbers of people with risks of overloading the communication infrastructure, SSNs can be used to offload mobile data traffic from cellular networks. As an example, cooperative smartphones can wisely forward user's data *hop by hop* to the nearest Wi-Fi access point so that the amount of offloaded data can further increase.

Enabling Technologies

In this section, we provide a quick review of existing hardware and software technologies that might support the implementation of SSNs on today's smartphones.

Communication Technologies

Communication among smartphones can be supported by a plethora of short/medium range communication interfaces and technologies; in the following, we review the main features of existent or emerging wireless communication technologies available on today's smartphones, in order to discover the more suitable (if any) for the deployment of SSNs.

Bluetooth offers a short communication range and a rate that ranges from 1.2 Mb/s in the early releases (version 1.2) to a maximum of 24 Mb/s in the latest releases (version 4). Bluetooth requires a significant involvement by the user, since the interface should be activated, the recipient device should be discovered, and the content to be sent should be manually specified. The poor communication performance, the annoying procedures, and users' reluctance hindered the diffusion of Bluetooth-based SSNs.

Near field communication (NFC) technologies have been successfully proposed to support content sharing between smartphones. The very short communication range (on the order of a few centimeters) provides natural protection against malware diffusion. However, such strength is also a limitation, since the utilization of NFC-based SSNs is limited to a few specific scenarios such as delay-tolerant content sharing, in addition to the need to involve user supervision and management.

Recent smartphones are equipped with IEEE

		Network mode	User's involvement in configuration	Data rate	Communication range	Energy consumption	Number of devices	Multi-hop
Technologies	Bluetooth	Infrastructure-less	High	Medium (up to 24 Mb/s)	Low ~100 m	Low	Few	Yes
	NFC	Infrastructure-less	Low	Low (up to 424 kb/s)	Very low (up to 20 cm)	Low	Few	No
	Wi-Fi	Infrastructured	Low	High (up to 600 Mb/s for 802.11n with MIMO support)	Medium ~ 250 m outdoor	High	Many	Yes
		Infrastructure-less "ad hoc mode"	High				Few	Yes
		Infrastructure-less "direct mode"	Low				Few	No
	HSDPA HSUPA	Infrastructured	Low	Medium (up to 42.3 Mb/s)	High tens of kilometers	High	Many	No
	LTE	Infrastructured (may also be infrastructure-less in Release 12 — D2D)	Low	High (up to 500 Mb/s in the Advanced version)	High tens of kilometers	High	Many	No

Table 1. Enabling technologies for SSNs.

802.11 interface (Wi-Fi) supporting high transmission rate up to 600 Mb/s (IEEE 802.11n). However, most existing mobile operating systems (MOSs) do not allow configuration of the Wi-Fi interface in ad hoc mode, unless breaking some safety procedures and creating a super-user account, which is quite far from being practical for common users. Hence, direct communication between smartphones is practically infeasible.

Wi-Fi Direct has recently been standardized to allow communication between enabled devices without any infrastructure. The devices activate a negotiation procedure at their first connection to determine which one shall act as an access point, while the other devices can connect to it by using, de facto, the Wi-Fi infrastructured mode. However, despite the benefits in terms of increased security compared to the ad hoc mode, Wi-Fi Direct causes an unfair workload distribution since the smartphones acting as access points will consume more resources. Furthermore, Wi-Fi Direct does not support multihop communication, and consequently, communications are limited to terminals within reciprocal communication range.

A further opportunity to support SSNs is given by LTE and also high-speed downlink/uplink packet access (HSDPA/HSUPA). Even if these technologies are mainly conceived to provide the devices connectivity toward a network infrastructure, it is expected that LTE in Third Generation Partnership Project (3GPP) Release 12 will support device-to-device (D2D) communications mostly based on local, opportunistic, and single-hop data exchange.

In Table 1, we briefly report the main characteristics of the wireless technologies discussed so far.

Software Support

While the characteristics of communication at the lower layers (medium access control/physical, MAC/PHY) depend on the wireless technology in use, the network functionalities are provided by the Mobile Operating System (MOS). Two main functionalities are fundamental to implementing an SSN: the possibility of turning a smartphone into a network gateway and the routing capabilities. Nowadays, most MOSs (i.e., Android, IOS, Windows Phone) make "tethering" modules available, which enable a smartphone to provide mobile Internet connectivity through its short-range wireless interfaces (e.g., Wi-Fi, Bluetooth). However, the enabling/disabling of tethering modules must be performed manually by the user. Routing capabilities on Wi-Fi networks are not provided by most popular MOSs for security reasons, even if the network interface card (NIC) usually supports ad hoc communication. Existing software can be categorized into two approaches [3]: delay-tolerant network (DTN)-based [4] or mobile ad hoc network (MANET)-based [5, 6]. The first approach allows the problem of intermittent connectivity caused by end-user mobility to be faced, although it poses severe challenges in terms of performance and battery consumption [4]. The second approach relies on traditional routing protocols used over generic MANET; we cite experimental studies of multihop SSNs using Optimized Link State Routing (OLSR) [2] and Ad Hoc On-Demand Distance Vector (AODV) [5] protocols. Since most of the MOSs do not allow the interception of IP packets or modification of the routing tables at the kernel layer, these implementations run in user space [5], thus introducing additional overhead

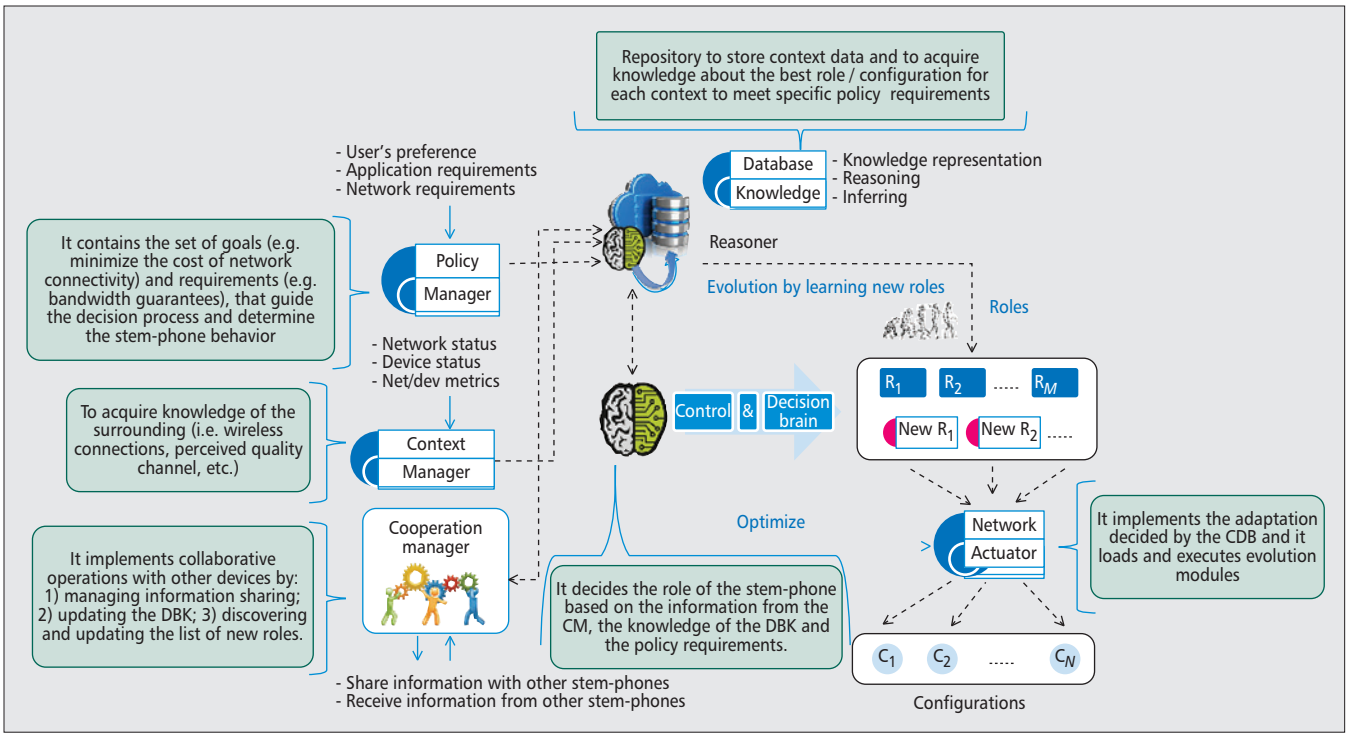


Figure 2. The generic architecture of a Stem-Phone.

in terms of power consumption and data transfer. At the same time, since the routing policy is performed on the basis of the route length only, energy consumption can be unbalanced or lead to suboptimal SSN performance.

Challenges in the Implementation of SSNs

From all the issues reported above, we argue that the main challenge to the implementation of SSNs is the lack of adequate software support for network creation and management. In this article we focus on three limitations of existing software architectures:

•*No autonomous device configuration capabilities*. In most of today's smartphones, configuration and setup of network functionalities must be carried out by users manually, like pairing operations with other devices/networks, network formation, and adjustment of transmitting parameters. This approach is not scalable and thus not suitable for scalable deployments (e.g., the pervasive access scenario) or dynamic environments (e.g., the emergency scenario).

•*No cooperative network management capabilities*. Although computation and communication capabilities of smartphones are continuously improving, their performance is still not comparable with that of dedicated network equipment. For instance, in [1], the authors have compared the performance of a SO-HO Wi-Fi network router to a smartphone provided with mesh routing capabilities, and have found that in this latter case power consumption constitutes a severe concern, since lookup operations are highly demanding for the smartphone's CPU. Vice versa, energy efficiency of a single device can be improved when smartphones cooperate to share the effort of network management, by dynamically deciding which role to take (e.g., router/gateway) on the basis of their actual resources.

•*No network self-organization capabilities*. SSNs are intrinsically dynamic environments, due to the end users' mobility and the variable traffic loads produced by the mobile applications. Self-organizing principles are required to manage the network and guarantee continuity of service in the face of dynamic and unforeseen events. Moreover, in several scenarios (e.g., post-disaster recovery), the goal of the SSN becomes to maximize the operativeness of the network, considered as a single entity, rather than the performance of a single component. Achieving such distributed intelligence requires cooperation, autonomous sensing, and decision making capabilities, which of course cannot be managed in the case of human control.

STEM-Net: A Framework to Support the Implementation of SSNs

In this section we describe the STEM-Net framework, originally introduced in [2] for a Smart Cities environment, and we show how the STEM-Net paradigm can overcome the previous challenges becoming a viable and effective solution also in the SSN context.

The logical architecture of a generic end-user smartphone is illustrated in Fig. 2. We refer to the *stemness* property of a smartphone as the ability to perform a protocol reconfiguration achieved throughout the combination of built-in features and algorithmic solutions. From here

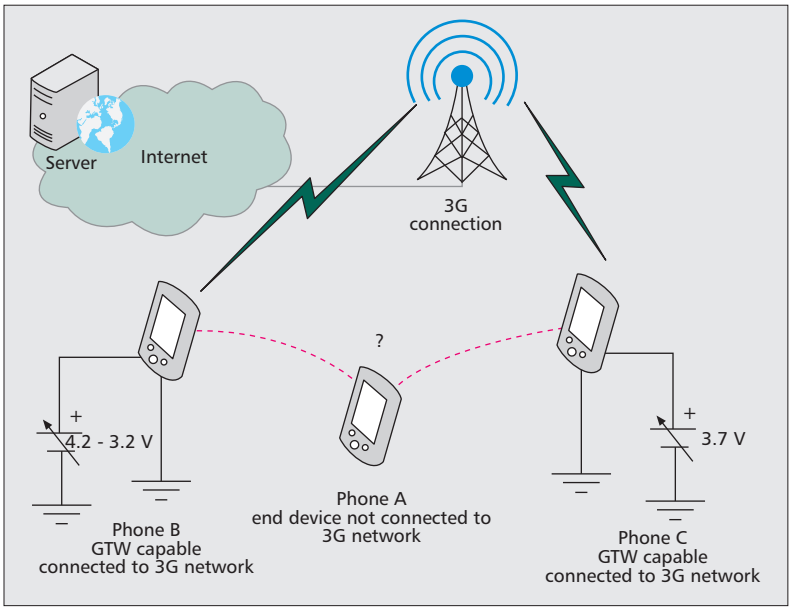


Figure 3. The testbed setup.

on, we use the term Stem-Phone to indicate this novel family of software enhanced smartphones. We further note that in our view, Stem-Phones may be a limited fraction of the existing smartphones but can still provide useful services to neighboring legacy phones.

Each Stem-Phone participating in the SSN setup can play a given set of roles according to the network capabilities/functionalities supported by the specific device. The basic set of roles of each Stem-Phone include the ability to produce/receive data, forward the traffic of other terminals, and to act as a gateway providing access to global network resources to other terminals.

The set of roles played by each Stem-Phone can vary in accordance with its own built-in characteristics (i.e., hardware features or physical constraints). For example, the gateway role could require the simultaneous use of different communication technologies to connect other nodes (e.g., Wi-Fi Direct, Bluetooth) and access global network resources (e.g., LTE, HSDPA). In addition, a Stem-Phone could also play new roles that can be dynamically configured and "learned" from other Stem-Phones by relying on cooperation with them. This might be the case, for instance, of a Stem-Phone that upgrades its software, downloaded from its neighbors, in order to gain the ability to serve as an access point. Moreover, each role is mapped to a specific network configuration, and the possibility is foreseen that a Stem-Phone may change its configuration over time for self-optimization purposes. This might be the case, for instance, of a Stem-Phone configured as a Wi-Fi router that dynamically adjusts its transmitting power level based on measured interference conditions.

Stem-Net For SSNs: The Gateway Election Case

In this section we show how the STEM-Net framework can usefully be applied to handle autonomous configuration tasks by focusing on a basic SSN issue: dynamic gateway election.

Let us consider a set of smartphones that cooperate to build an SSN. Without loss of generality, we suppose that multihop communication is supported by following the approach proposed in [1]. Wi-Fi cards are configured in ad hoc mode, and the OLSR protocol is deployed. Some of the smartphones should be elected to play the role of gateway, which consists in forwarding the packets coming from the SSN on the Wi-Fi interface toward the Internet backbone by using the 3G interface. This role requires the fulfillment of several requirements:

- Acting as a gateway implies high power consumption; therefore, the gateway should be chosen among those Stem-Phones with sufficient residual energy.
- The connection between the gateway and the cellular backbone should offer adequate throughput to sustain the traffic generated in the SSN.
- Since the throughput on Wi-Fi multihop communication roughly decreases with the number of hops, the gateway should be located in a central position with respect to the other terminals in the SSN.
- The Wi-Fi network near the gateway should be as uncongested as possible.

To address this problem, a simple but effective spontaneous gateway election procedure has been proposed in [7]. Since we only aim to show how this election procedure is functional to SSNs, we briefly summarize our strategy in the following by referring the reader to the work in [7] for further details. Once an SSN is created, a gateway is randomly selected; after this initialization phase, the gateway role is passed to the most suitable node by following a *stimulus-response* model. Each node monitors the following parameters:

- Its own residual energy
- The congestion level experienced at the Wi-Fi interface
- The average distance from other nodes in the SSN

The gateway g also computes the congestion experienced at the cellular interface. All the monitored parameters are periodically exchanged between the nodes. The exact computation of the relevant parameters is given in Table 2.

The node acting as gateway constantly evaluates its attitude to keep its role and quantifies it through a specific metric called the stimulus metric (SM) [7],

$$SM(g) = G(\textit{residual energy}, \textit{congestion on cellular interface}).$$

Here, G represents a suitable function, which is given in Table 2. Periodically, the gateway in charge broadcasts its SM value by starting a *gateway handover procedure*; upon receiving an SM message, each Stem-Phone i evaluates its capability to take over the gateway role and summarizes it in a new metric called the threshold metric (TM), defined as follows [7]:

$$TM(i) = F(\textit{residual energy}, \textit{congestion on WiFi interface}, \textit{average distance from neighbors})$$

Section A		
Relevant parameters estimation	**Parameters**	**Computation method**
	$E(i)$: residual energy at node i	Each operating system offers specific routines.
	$C_{cell}(i)$: congestion on the cellular interface of node i	The interface driver offers an estimation of total transmitted TX_{Cell}, received RX_{Cell}, dropped $DROP_{Cell}$ and corrupted ERR_{Cell} packets: $C_{cell} = \frac{DROP_{cell} + ERR_{cell}}{TX_{cell} + RX_{cell}}$
	$C_{Wi\text{-}Fi}(i)$: congestion on the Wi-Fi interface of node i	The interface driver offers an estimation of total transmitted TX_{WiFi}, received RX_{WiFi}, dropped $DROP_{WiFi}$ and corrupted ERR_{WiFi} packets: $C_{WiFi} = \frac{DROP_{WiFi} + ERR_{WiFi}}{TX_{WiFi} + RX_{WiFi}}$
	$D(i)$: average distance from neighbors of node i	The OLSR routing tables include the hop distance from other nodes in the SSN. $D(i)$ is the average of those hop distances.
Section B		
Stimulus and threshold computation	**Functions**	**Computation method**
	$G(\ldots)$: stimulus at node g	$SM(g) = a(1 - E(g)) + \beta(C_{cell}(g))$ where the smoothing factors α and β are chosen to respect the relation $\alpha + \beta = 1$
	$F(\ldots)$: threshold at node i	$TM(i) = 1 - \left[\gamma \frac{E(i)}{E_{\max}} + \delta\left(1 - \frac{C_{WiFi}(i)}{C_{WiFi}^{\max}}\right) + \varphi\left(1 - \frac{D(i)}{D_{\max}}\right)\right]$ where the smoothing factors γ, σ, and ϕ are chosen to respect the relation $\gamma + \sigma + \phi = 1$

Table 2. Parameters estimation and stimulus-threshold computation.

Here, F is a suitable function, given in Table 2. If the stimulus $SM(g)$ perceived by a node i exceeds its threshold $TM(i)$, node i assumes the role of gateway and announces its decision to the SSN.

We would like to remark here that the described gateway election procedure can easily be supported by the proposed Stem-Phone architecture (Fig. 2); in particular, the *context manager* estimates the most relevant parameters (RPs) (residual energy, congestion on Wi-Fi and cellular interfaces, average distance), the *cooperation manager* takes care of the RPs exchanges between smartphones, the *knowledge database* stores the RPs concerning both the smartphone itself and neighboring ones. The *policy manager* translates the user preferences such as his/her reluctance to share the residual charge and their acceptable performance. Finally, the *control & decision brain* computes the metrics $SM(g)$ and $TM(i)$, and makes decisions.

A Proof-of-Concept Implementation

The effectiveness of gateway election has already been shown in [7]. Here, we are interested in providing some brief guidelines on the feasible implementation of the election process on commonly available smartphones. As a hardware platform we used a Samsung Galaxy S model, which can be considered representative of a wide range of user devices on the market.

On top of the hardware platforms we have developed a software suite, which implements the gateway election procedure described above. The software suite includes two classes of programs that implement:

- The basic communication functions
- The components of the Stem-Phone architecture

The basic communication over multihop paths is supported by using the OLSR daemon (www.olsr.org).

The components of the Stem-Phone architecture have been implemented as follows:

- The context manager consists of a series of self-developed scripts that also leverage some routines offered by the Android operating system. Specifically, Table 2 summarizes how RPs are computed.
- The cooperation manager has been developed by extending the OLSR protocol with two custom message types:
  - –STEM HELLO MESSAGE to disseminate *RPs* computed by each Stem-Phone
  - –STEM GATEWAY MESSAGE to start a gateway election procedure.

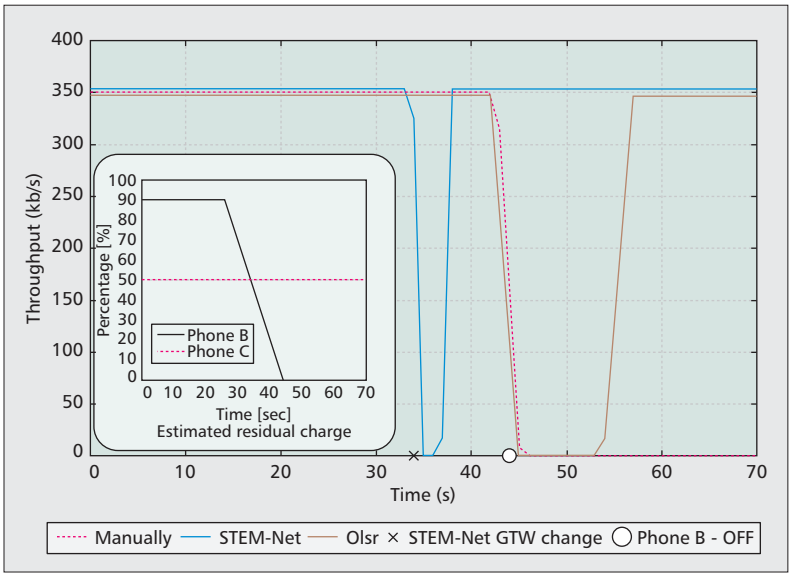


Figure 4. Throughput experienced by phone A using different residual charge for both phones B and C.

- The knowledge database is implemented through a set of files in the *~/proc/Stem/* directory. Specifically, a file named *~/proc/Stem/context.local* stores the information relevant to the node itself, while the file *~/proc/Stem/context.extra* is filled with the information received by surrounding nodes.
- The control & decision brain functionalities have been supported by implementing a dedicated C program that, starting from the information stored in the ***knowledge database***, computes and compares the stimulus and threshold according to the algorithm described earlier.

The realized prototypes have been tested in the deployment shown in Fig. 3, where three Stem-Phones constitute a small SSN. One of the three phones (phone A) is unable to directly transmit data to the 3G network. This condition has been forced by instructing the Android OS to disable data communication through the 3G interface. Hence, phone A should forward its data to one of the other two phones (B or C) that instead have their 3G connection active. Data produced by phone A toward the Internet consists of a constant bit rate flow of 350 kb/s, generated using custom software.

To control the residual charge available to phones B and C, their batteries have been replaced with wooden ones, and the phones have been powered through a tunable power supplier. Commonly available phones estimate the state of charge (SOC) of their battery by probing the voltage provided by the latter. Thus, in our setup we forced the phones to change their estimation of the residual charge by varying the voltage value throughout a tunable power supply. In our test campaign the voltage provided to phone C was kept constant to 3.7 V to emulate a residual battery charge of about 50 percent of the nominal value; on the contrary, the voltage provided to phone B was varied in [3.2–4.2 V] to emulate the residual amount of charge shown in Fig. 4.

Three different approaches to gateway election have been evaluated:

- Phone B is manually selected. This approach is equivalent to tethering operation, which is the only solution commercially available today.
- OLSR is installed on all the phones. In particular, phones B and C are configured as potential gateways and the choice of which to select is left to the routing protocol. This case represents what can be obtained through the framework presented in [1] or using similar MANET-based approaches.
- The STEM-Net architecture is deployed on the phones through the software modules described earlier.

Figure 4 reports the throughput achieved by phone A when varying the gateway selection method and the residual charge at phones B and C. The worst performance is achieved when the gateway is manually chosen. In this case, as soon as the residual charge at phone B approaches zero, the latter switches off its communication interfaces and stops forwarding data coming from phone A; hence, the communication is interrupted.

The performance improves when the choice of the gateway is carried out by using OLSR. In this case Phone B is selected as gateway as long as its available charge goes to zero. At this instant, the communication path is broken and OLSR looks for a new available gateway. After about 10 seconds Phone C is selected as new gateway. This behavior constitutes an improvement compared to the first case (manual selection) nonetheless it presents some drawbacks:

- It leads to an uneven power consumption between the available gateways
- The search for a new gateway is started only after Phone B became inoperative and following this late reaction a service disruption of about (10s) is experienced
- The owner of Phone B may be annoyed to have his/her phone inoperative due to the lack of charge

In the STEM-Net framework the gateway election procedure is carried out also considering the residual charge, so the role of the gateway is always assumed by the node with the maximum residual charge. As soon as the residual charge of phone B falls below that of phone C, the role of gateway is passed from the former to the latter. Since the change of role is done while phone B is still operative, the handover procedures are much faster than in the OLSR case. Finally, the STEM-Net approach does not starve phone B; thus, it provides a better experience to the end user.

This simple experiment confirms the viability and effectiveness of the STEM-Net framework on commonly available devices and paves the way for additional experimentation for upcoming SSNs.

Lessons Learned and Open Issues

Based on the experimental results, we can conclude that STEM-Net represents a viable solution to face the challenges to SSN

implementation introduced earlier. Indeed, cooperative network management policies can be implemented in a straightforward way through the concepts of roles and node mutation, while network adaptiveness and distributed self-organization capabilities can be achieved through the proper modeling of stimulus/threshold functions, which are intrinsically scenario-dependent. However, several practical issues regarding autonomous device configuration capabilities have emerged from the testbed implementation, based on which we can argue that a completely user-agnostic SSN deployment model might not be 100 percent feasible on today's smartphones. Skipping the technical details, most of such issues derive by the poor support provided by the MOS at the application programming interface level, so several low-level network functionalities (e.g., multihop routing) cannot be implemented without breaking some safety mechanisms of the original software equipment, which is clearly not a solution.

Finally, we conclude the article by highlighting the existence of additional open issues that need to be addressed for practical deployment of SSNs. Security is a major concern while designing and developing spontaneous network based on mobile devices [8]. A secure self-configured protocol is required for user authentication, validation, and data transfer. Moreover, novel and robust reputation mechanisms are required to identify user misbehavior within different communities [9]. Connected to the security issues, novel cooperation models and utility functions have to be designed to encourage users to share their resources with the certainty of obtaining future benefits [10]. In this context, smart mechanisms based on local pricing strategies could be used to control and prevent network congestion by prolonging the overall lifetime of the SSN, considered again as a single entity. We plan to address these open issues in future work.

Conclusion

In this article we have investigated the potential of smartphone spontaneous networks by analyzing the main issues and challenges from the perspective of minimizing users' involvement in network setup and maintenance. We believe that the implementation of an actual network *spontaneity* model, based on *software* intelligence local at each smartphone, could favor the diffusion of SSNs. To this purpose, we have introduced the STEM-Net framework, and presented results of a small-scale testbed in which gateway nodes are dynamically selected.

References

[1] A. Iera *et al.*, "Making A Meshrouter/Gateway from a Smartphone: Is That a Practical Solution?," *Ad Hoc Networks*, vol. 9, issue 8, 2011, pp. 1414–29.

[2] G. Aloi *et al.*, "STEM-Net: An Evolutionary Network Architecture for Smart and Sustainable Cities," *Trans. Emerging Telecommun. Technologies*, Special Issue: Smart Cities — Trends & Technologies, vol. 25, issue 1, Jan. 2014, pp. 21–40.

[3] 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. 56–63.

[4] H. Ntareme, M. Zenna, and B. Pehrson, "Delay Tolerant Network on Smartphones: Applications for Communication Challenged Areas," *Proc. ACM ExtremeCom*, Galapagos Islands, Ecuador, 2011.

[5] T. Zhuang, P. Bskett, and Y. Shang, "Managing Ad Hoc Networks of Smartphone," *IJIET Int'l. J. Info. and Education Technology*, vol. 3, issue 5, 2013, pp. 540–46.

[6] P. Mitra and C. Poellabauer, "Emergency Response in Smartphone-Based Mobile Ad Hoc Networks," *Proc. IEEE ICC*, Ottawa, Canada, 2012.

[7] M. Di Felice *et al.*, "Smartphones Like Stem Cells: Cooperation and Evolution for Emergency Communication in Post-Disaster Scenarios," *Proc. IEEE BlackSeaCom*, Batumi, Georgia, 2013.

[8] R. L. Gilaberte *et al.*, "A Secure Protocol for Spontaneous Wireless Ad Hoc Networks Creation," *IEEE Trans. Parallel Distrib. Sys.*, vol. 24, issue 4, 2013, pp. 629–41.

[9] R. Sofia *et al.*, "Moving toward a Socially-Driven Internet Architectural Design," *Computer Commun. Rev.*, vol. 42, issue 3, 2012, pp. 39–46.

[10] P. Pace and G. Aloi, "Managing and Deploying Pervasive Wireless Internet Access through Attractive Connection Sharing and Reselling Mechanisms," *J. Networks*, vol. 8, issue 2, 2013, pp. 351–64.

Smart mechanisms based on local pricing strategies could be used to control and prevent network congestion by prolonging the overall lifetime of the SSN, considered again as a single entity. We plan to address these open issues as future works.

Biographies

Gianluca Aloi [S'99, M'02] (aloi@dimes.unical.it) received his M.S. degree in computer science in 1999, and his Ph.D. degree in systems engineering and computer science in 2003. He joined the University of Calabria in 2004, where currently he is an assistant professor in the Department of Informatics, Modeling, Electronics and Systems (DIMES). His research interests include spontaneous and reconfigurable wireless networks, cognitive networks, resource provisioning and sharing, software defined radio systems, localization systems, and satellite communications.

Marco Di Felice (difelice@cs.unibo.it) received his Ph.D degree in computer science in 2008 from the University of Bologna, Italy. In 2008 and 2010, he was a visiting researcher at the Georgia Institute of Technology, Atlanta, and Northeastern University, Boston, Massachusetts. Since April 2012, he has been an assistant professor in computer science at the University of Bologna. He has authored more than 60 papers on wireless and mobile systems and applications.

Valeria Loscrì (valeria.loscri@inria.fr) has been a permanent researcher at Inria Lille — Nord Europe (FUN Team) since October 2013. She got her Master's degree in computer science and Ph.D. in systems engineering and computer science in 2003 and 2007, respectively, at the University of Calabria. In 2006 she was a visiting researcher at Rice University, Houston, Texas. She has authored more than 60 publications in journal, conferences, workshops, and book chapters. Her research interests focus on self-organizing systems, robotics-networks, and nanocommunications.

Pasquale Pace [S'02, M'05] (ppace@dimes.unical.it) received his Ph.D. in information engineering in 2005 from the UNICAL, Italy. In 2005 and 2006, he was a visiting researcher at the CCSR, Surrey, United Kingdom, and at the Georgia Institute of Technology. He is currently an assistant professor in telecommunications at the UNICAL. He has authored more than 60 papers in international publications. His research interests include cognitive networks, sensor and self-organized networks, and cost and business models for resource sharing.

Giuseppe Ruggeri (giuseppe.ruggeri@unirc.it) received his Master's degree in electronics engineering in 1998 from the University of Catania, Italy. In 2002 he received his Ph.D. in electronics, computer science, and telecommunications engineering from the University of Palermo, Italy. He is currently an assistant professor in the Department of ICT, Infrastructures, and Sustainable Energy (DIIES) at the University "Mediterranea" of Reggio Calabria, Italy. His research interests focus on wireless networking, self organizing networks, and information-centric networking (ICN).

Content Dissemination in Vehicular Social Networks: Taxonomy and User Satisfaction

Farouk Mezghani, Riadh Dhaou, Michele Nogueira, and André-Luc Beylot

ABSTRACT

Social networking applications have gained huge popularity. With the widespread use of smart devices (e.g., on-board units, smartphones), these social networks are increasingly going mobile. As a result, a new trend of networking has emerged, referred to as vehicular social networking (VSN), which combines the wireless communications between vehicles with their social relationships. In a broader view, VSNs are social networks formed on roadways by users who have social relationships, interactions, and common interests. The exploitation of vehicular users' social properties provides better networking and social support to innovative applications and services. This article overviews recent achievements in VSN by providing an organized view of existing approaches. Its contribution lies in a taxonomy for content dissemination approaches in the context of VSN. Also, a framework is outlined to tackle a major new challenge: supporting user satisfaction. Finally, this article emphasizes open research and future trends.

INTRODUCTION

Social networks have gained significant attention in the research and industrial communities. With the rapid evolution of the Internet, online social networks (OSNs) such as Facebook appeared as the first form of social networking, but have been limited to online activity. With the advent of wireless mobile devices (e.g., smartphones, onboard units — OBUs) that have the capability to detect proximity to other users, and communicate and share data with them, various types of networks are emerging as new paradigms to exploit social properties of mobile nodes such as vehicular social networks (VSNs), mobile social networks (MSNs), and delay-tolerant networks (DTNs).

VSNs [1] enable drivers and passengers who usually travel every day between home, office, and points of interest to socialize and exchange information with other commuters on the road. These commuters may perceive the traffic situation and share the driving experience (e.g., road hazards and traffic jams) in order to enhance traffic management. Furthermore, they can support the exchange of useful information for commuter entertainment (e.g., gas prices and video news). Due to the resource constraints of mobile devices and communication networks, content dissemination in VSNs presents several challenges. For example, due to the high dynamism of the network, it is hard to understand the social properties of the nodes and how to take advantage of users' behavior to improve the performance of the network in terms of content dissemination [2]. Communication links between vehicles might remain active only for short periods of time because of the high mobility of vehicles.

In the last decade, researchers have begun to address these issues. However, the literature on VSN lacks work that can present the state-of-the-art challenges in content dissemination. Furthermore, the literature lacks work pointing out the main characteristics of existing approaches for content dissemination and outlining open issues. Hence, this article:

- Reviews recent achievements related to content dissemination in VSNs
- Provides an organizing view of existing approaches by clearly pointing out their advantages and constraints
- Facilitates deep study of advances in the state of the art

Furthermore, it has been observed that interesting approaches exploit classic metrics such as delivery delay and delivery ratio to support the content dissemination schemes. Although those solutions represent improved results, another important feature has been left behind: *user satisfaction*, a quantitative metric that computes how well users are satisfied. It calculates, through a function of users' interests, the benefit (gain) of users in receiving the content. In VSN content dissemination is interest-driven; hence, it is necessary to provide a scheme that maximally satisfies users' interests (i.e., maximize the total data utility by sending the appropriate content objects that match user interests). This work presents a framework to measure and maximize the satisfaction of users' interests.

Farouk Mezghani, Riadh Dhaou, and André-Luc Beylot are with the National Polytechnics Institute of Toulouse.

Michele Nogueira is with the Federal University of Paraná.

0163-6804/14/$25.00 © 2014 IEEE

The contributions of this article are first to provide the reader a comprehensive view of content dissemination in VSNs, paying particular attention to the different ongoing solutions, pointing out their limits and advantages. It allows emphasis to be placed on open research issues that the research community is called to address. Second, for tackling one of the main challenges, the lack of user satisfaction, a framework that targets maximally satisfying users' interests is proposed.

This article proceeds as follows. We provide a background on VSNs. Then a taxonomy for the VSN content dissemination approaches is presented. Next, we present a framework as a proof for the need to consider the user satisfaction feature in content dissemination schemes. Finally, we draw important directions for future studies.

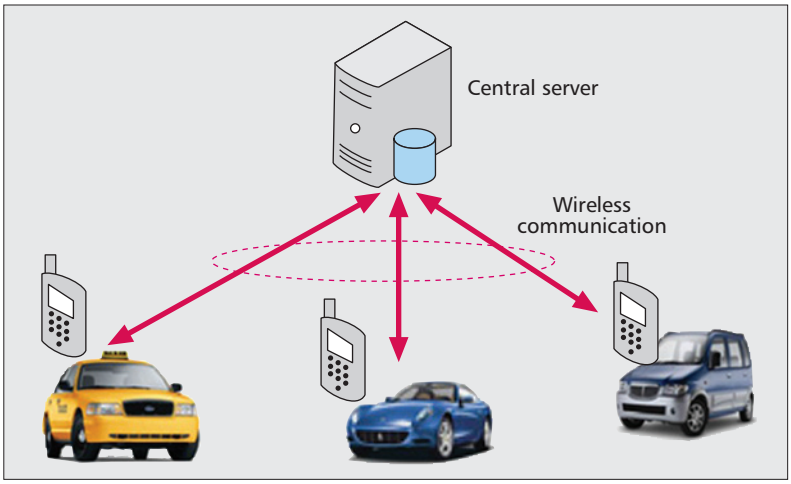


Figure 1. Centralized VSN.

Background

Difference between VSNs and MSNs

An MSN is defined as a social network formed by mobile handsets carried by humans. When these participants are moving in vehicles (drivers, passengers), they form a so-called VSN. Furthermore, in VSNs other equipment such as OBUs can be used on board vehicles.

From the social networking point of view, social relationships between users in MSNs are considered stronger than the ones in VSNs. Indeed, in MSNs, users have a very high probability of meeting each day and for enough time to exchange a lot of information (e.g., colleagues at a university, employees at work). On the other hand, VSN users do not have the same probability to meet even if they share the same destination. This is because of the highly dynamic topology and high speed of vehicles.

From a communication point of view, vehicles always move at high speed (except at intersections) in very large areas, while in MSNs, users are usually walking in a confined area (e.g., within a campus). Thus, contact frequency between the same users is very different between VSNs and MSNs. Additionally, contact duration between users is much longer in MSNs than in VSNs. Indeed, it is on the order of minutes for MSNs instead of seconds in VSNs. For instance, considering a communication range of 200 m (direct WiFi or IEEE 802.11p), the minimal contact duration for moving humans is equal to 2 min 22 s. For moving vehicles this duration is about 15 s when speed is equal to 50 km/h [3, 4]. The limitation of duration shows how difficult exchanging data is in a vehicular environment, especially for contents of large size. Even so, in MSNs users connect for enough time; they can exchange many objects, even of large size. Furthermore, regarding the communication constraints, only MSN users have energy constraints since VSN users can benefit from the power resources in vehicles.

These differences show the complexity of VSNs in relation to MSNs. Therefore, content dissemination protocols proposed and used in MSNs are either incompatible or not efficient for VSNs.

Vehicular Social Networks

The main components of a VSN can be defined as follows:

- Participants
- Mobile device
- Network infrastructure

In vehicular networks, not only can drivers participate, VSNs, but also passengers in the vehicles. Each user can be either a content provider or a consumer. Mobile devices can be integrated in the vehicles (e.g., OBUs) or carried by users (e.g., smartphones), enabling the detection of proximity to other users' devices, and communicating and sharing content. Network infrastructure, such as cellular networks and RSUs (roadside units), is usually used in VSNs for centralized applications.

According to the way users are able to access and deliver data objects, VSNs can be broadly classified into three types of architecture: centralized, decentralized, and hybrid. The centralized architecture of VSNs, as shown in Fig. 1, works under the assumption that users must continuously access a centralized server, which coordinates and manages their interactions with other users, even when the vehicles are physically close. In such an architecture, there is no direct interaction between vehicles. Vehicles interact directly with the infrastructure of support, mainly RSUs; this communication is referred to as vehicular-to-infrastructure (V2I) communication. Vehicles communicate indirectly by means of the RSUs.

The social relations and personal profile of each user may be deemed to be relatively stable (i.e., preserved in the central server) unless the VSN users update their profiles and interests or their friendship ties over time.

A decentralized VSN, as shown in Fig. 2, includes social networks only enabled by opportunistic vehicle-to-vehicle (V2V) contact. This concept leads to benefits from both physical and virtual communication. Content dissemination in infrastructureless VSNs is a challenging task since it requires users to collaborate without the aid of a central entity. When different nodes on the road, sharing common interests, are in proximity, they can establish a temporary community

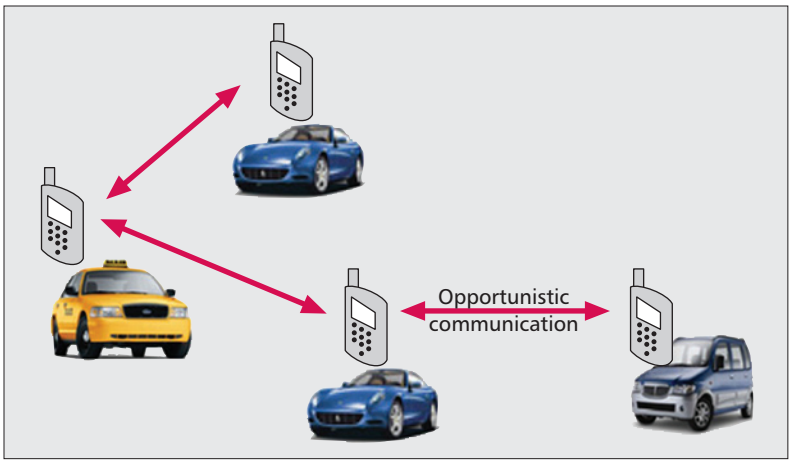


Figure 2. Decentralized VSN.

by self-organization and share data objects. Then the community can be broken up once they complete the dissemination process.

A hybrid architecture comprises both V2I and V2V communications. RSUs that are specifically designed for vehicular networks are not deployed often due to the excessive cost of their deployment. Thus, nowadays, several vehicular network applications are based on cellular networks due to the lack of RSUs. Beyond on-road safety, many services that are user-oriented are emerging for vehicular communications systems. In particular, social networking can be useful and helpful for vehicle traffic efficiency and infotainment. Two major types of VSN applications are considered:

- *Traffic management*
- *Entertainment*

Traffic management applications such as WAZE [5] and NaviTweet [6] provide user information about traffic collected in real time such as traffic jams and approaching police. Entertainment applications may be useful for drivers and passengers, especially on a long trip. For example, a voice-based application [7] enables drivers sharing common interests and driving in the same roads to interact using voice messages.

Content Dissemination in Vehicular Social Networks

This section highlights the evolution of content dissemination approaches in VSNs. Figure 3 presents the proposed classification. It consists of three main categories:

- Information processing
- Content delivery
- Performance

From the perspective of information processing, which represents the way VSN dissemination approaches treat the information, VSN dissemination is classified into three parts. Information relevance is ignored (i.e., considered as a black box), estimated, or taken from users' (personal) preferences. From the content delivery perspective, VSN dissemination schemes are categorized as utility-based or blind-based. And from the perspective of performance, VSN content dissemination approaches follow three main features: delivery delay, delivery ratio, and used bandwidth. The following subsections detail each category providing the state-of-the-art VSN content dissemination approaches.

Note that an approach is not exclusive to a single category.

Information Processing

In the first category, the classification is based on the way VSN dissemination approaches treat the different content objects. It can be classified into three types: information as a black box, estimating the information relevance for the user, or employing user preferences.

Information as a Black Box — Most previous empirical work [8–10] in the literature does not consider the subject of different content in VSNs. Information relevance for users is ignored, characterized only by its general features such as size and lifetime. Therefore, when delivering content, all vehicles are considered as targets. VSN applications are designed for commuters' comfort and entertainment; thus, nowadays users prefer to get few contents matching more with their interests and bringing more benefits than receiving several contents in which they are hardly interested. It is important to emphasize that dissemination in VSNs is characterized by information content. Different contents may correspond to different user interests.

Estimating Information Relevance — In [11], the authors propose a new technique to estimate the relevance of data for the drivers. Their solution identifies and classifies the information type to the vehicle in order to estimate the information relevance, then deciding whether to inform the driver or share the information further. The advantage of approaches that estimate information relevance is avoiding the exchange of users' interests. The main concern behind these approaches is that information relevance is related to the direction of the vehicle and the type of information, and not to the vehicular user's personal interests (i.e., the probability of a vehicle encountering an event increases as the relevance of this event increases). Estimating the information relevance makes no sense for some types of information, such as music, which is related to the users' personal interests.

Considering User Preferences — A third class of work assumes that content is related to the personal profile of the user [12, 13]. Each user defines her/his personal interests in order to receive content in which s/he is interested.

This method is best to match the users' interests. The only disadvantage is that there is a need to exchange preferences before data delivery.

Content Delivery

Content delivery is a challenging issue in VSNs due to the dynamic topology and intermittent connectivity. The second group of the presented taxonomy aggregates content dissemination approaches addressing these issues. Content dis-

semination algorithms are classified in two general types: *blind delivery* and *utility-based* delivery.

Blind Delivery — Several content dissemination algorithms are broadcast-based, such as the work in [8, 9]. The underlying principle is to use social communications to enhance the network resources. Blind-delivery-based dissemination can be effective for network resources, but may not be effective for users.

For example, in blind-delivery-based approaches, social relationships are used in order to accelerate the dissemination process, finding the appropriate forwarding nodes/links to increase the delivery efficiency and reduce the delay. Moreover, it can limit the bandwidth utilization by minimizing the number of nodes responsible for forwarding. Interaction between nodes is done in only one step, which consists of exchanging the data (i.e., there is no neighbor discovery or interests exchange). In [8], the authors propose a scheme for fast forwarding. This scheme uses the social relation ties between vehicles to choose the most appropriate data relays. Blind delivery can be efficient in terms of network resources, but might not be efficient for commuters. The main drawback of blind delivery approaches is the ignorance of users' preferences. In reality, social relationships are achieved between nodes that share common interests. Therefore, another important criterion that should be considered in the dissemination process is the users' interests.

Utility-based Delivery — In recent works [6, 12, 14], the authors present the problem of dissemination in VSNs from a different perspective. Even though content dissemination algorithms can deliver content rapidly to as many users as possible, many of these forwarded objects might not be useful for users, and they are ignored. For example when receiving an object, usually the user, according to the topic, decides to use or ignore the object. Social interactions are used to share content between users with common interests; thus, it is likely that the user shares the information in which s/he is interested. The probability of sharing information of less interest is low. This method may allow users to receive data in which they are interested while reducing the reception of uninteresting objects. For example, in [1] the authors proposed a VSN approach that enables commuters on the same road at the same time to communicate through voice messages about a specific topic. This proposal only considers one type of service, while different types of services need to be accommodated by VSNs. Therefore, there is a need for a multiservice dissemination scheme that efficiently explores users' interests.

Performance

The third group of the presented taxonomy comprehends VSN dissemination approaches that improve content dissemination performance based on a feature. Those approaches are classified into three groups, according to the employed performance metrics *delivery delay*, *delivery ratio*, and *bandwidth usage*.

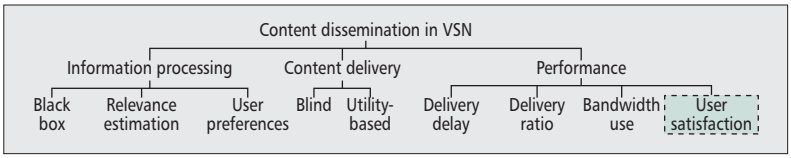


Figure 3. Taxonomy for content dissemination in VSNs.

Delivery Delay — This refers to the delay for a message to be received at the destination. The delivery delay is the major constraint for some applications, such as safety, traffic, and information, that has a short lifetime. The distributed information has a short time to live, and thus should be delivered to the destinations before it expires. For example, the work in [8] proposes a scheme based on social ties between vehicles for fast forwarding. However, delay-tolerant applications such as entertainment applications have fewer constraints on the delivery delay. For instance, information about a gas station or music has an unlimited or long lifetime; therefore, a short delivery delay is appreciated but not mandatory.

Delivery Ratio — This represents the ratio of data objects successfully delivered to destinations. The main concern behind this approach is that it considers all vehicles as destinations. For example, in [9] the authors propose a VSN scheme for mobile advertising that targets improving the coverage and intensity of advertising. However, in VSN entertainment applications, destinations refer to the nodes that are interested in a given data object. In this case, information is successfully delivered to nodes that are interested in it. Objects delivered to other nodes are ignored.

Bandwidth Usage — This refers to the use of network resources [10]. The use of bandwidth represents a major constraint: dissemination algorithms need to efficiently use network resources, not overload the system. Especially in a dense network, an excessive exchange of data and/or signaling may overload the network.

Discussion

The observation of the proposed taxonomy leads to different features that can improve content dissemination approaches. We have observed that existing works lack another important feature that evaluates the satisfaction of users. To meet this need, this article proposes adding a new feature, labeled *user satisfaction*, in the performance group shown in Fig. 3. The main goal is to maximize the total data utility by sending the appropriate content objects that match user interests. Indeed, users have different interests; thus, an efficient dissemination algorithm has to reduce the reception of useless objects. Moreover, a user has heterogeneous preferences; for example, a vehicle may be interested in traffic information more than a gas station notification. Therefore, the dissemination protocol should be based on an efficient method that can increase the satisfaction of users by delivering the appropriate objects. This feature may not be important in safety applications because these

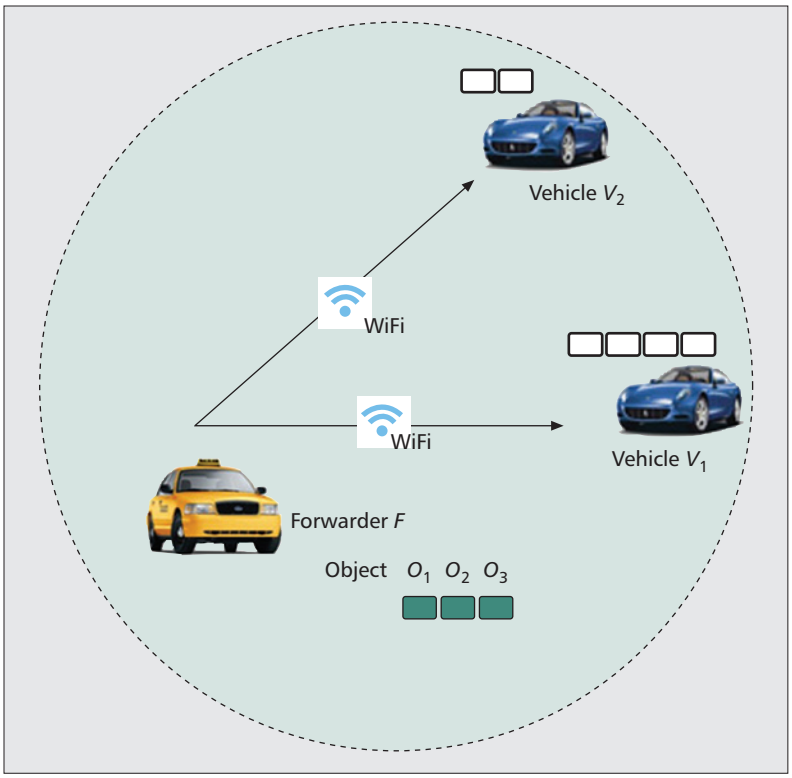


Figure 4. Simple scenario with different object interests and contact durations.

applications aim to increase the security for the drivers and do not aim to increase their satisfaction.

User Satisfaction Framework

VSNs constitute an environment where a large amount of heterogeneous contents are being generated every day. Users are seldom interested in all these contents; they only want specific useful information. Nowadays, with the growing popularity of personalizing applications, customers prefer to get content based on their personal interests. Moreover, connections between vehicles in a VSN occur only during a very short period, allowing users to exchange a limited volume of data. Therefore, there is an increasing demand for efficient content dissemination in VSNs that considers the heterogeneous preferences of users and targets maximally satisfying user interests.

In order to tackle this challenge, we propose a framework based on two main goals: first, the user preferences should be taken into consideration to deliver objects interesting to them; second, heterogeneous user preferences should be taken into account to distribute the appropriate objects accordingly in short periods of time.

Consider the example shown in Fig. 4, where a forwarder F owns three objects (O_1, O_2, O_3) and meets neighboring vehicles, V_1 and V_2. Each user V_i ($i \in \{1, 2\}$) has interest $I_{V_i, O_j} \in [0..10]$ for object O_j ($j \in \{1, 2, 3\}$), as shown in Table 1. Assume that the contact duration is only long enough for vehicles V_1 and V_2 to receive 4 and 2 objects, respectively.

Forwarder F	Contact duration	O_1	O_2	O_3
V_1	$4T$	7	2	4
V_2	$2T$	5	9	9
sum I_{V_i,O_j}		12	11	13
		sum $I_{V_i} = 36$		

Table 1. Example of simple scenario.

We define the metric of *user satisfaction*, $U_{satisfaction}$, to determine how well users' interests are satisfied. It computes the utility gained (benefit) of the users after receiving the different content objects from F.

When vehicles meet opportunistically, an efficient content dissemination strategy addresses which objects to forward and how to schedule these objects to maximally satisfy user interests. Based on classical dissemination approaches, the forwarder F disseminates objects randomly. However, this dissemination strategy cannot maximize satisfaction. If we take into account heterogeneous user preferences, users can distribute objects efficiently to attain better satisfaction for neighboring users. In particular, based on local interests, F distributes the objects in the following order: [O_3, O_1, O_2], since the satisfaction that can be obtained is 13, 12, and 11, respectively. In this case, the global satisfaction, which represents the gained cumulative satisfaction, obtained after the broadcast of all the objects of the forwarder F is 0.75 ((13 + 12 + 2)/36).

Additionally, considering other parameters such as *contact duration* between vehicles can enhance this forwarding strategy and ensure a high level of user satisfaction. The link between F and V_1 is maintained for enough time (V_1 can receive all 3 objects as opposed to V_2, which can only receive 2 objects). Then, since object O_2 is more important to V_2 than object O_1, F can send the objects in the following order [O_3, O_2, O_1]. Thus, the global achieved satisfaction is: 0.86 ((13 + 11 + 7)/36).

Even though content dissemination approaches enable content to be delivered quickly to many users, they might not guarantee the satisfaction of users; hence, as this example shows, the need for a *user satisfaction* metric. This metric considers user preferences and allows the appropriate objects to be distributed efficiently.

Simulations are conducted to evaluate a scenario of 100 equipped vehicles with a transmission range of 200 m. One thousand objects, with a time to live set to 1 h, are generated in the beginning of simulations and distributed randomly over 10 users as initial data sources. For each user, a list of interests are associated using a uniform distribution. Simulations compare the following schemes:

- **Epidemic**: Each forwarder randomly schedules its set of data.
- **Local interest**: The forwarder sorts its data by their importance to the receivers (i.e., considering only the neighbors' interests).

- **Interest- and contact-duration-based**: Content objects' scheduling is based on both duration of nodes' contact and user interests.

Figure 5 shows the user satisfaction rate produced in the network over time. When all objects lifetime have expired (i.e., at t = 60 min), interest- and contact-duration-based satisfied 0.94 of user interests, while local interest and epidemic satisfied 0.845 and 0.747, respectively.

Challenges, Open Issues, and Future Directions

VSNs are a very recent research domain, and many challenges remain yet to be addressed. This section presents some future research directions in this field.

Selfish Users

Unlike the classic selfish behavior considered in different works, in the context of VSNs, two types of selfishness are considered: individual and social user selfishness. Individual selfishness refers to a node that is always looking out for its own interests. In contrast, from the social perspective, a selfish user is usually willing to cooperate with other users with whom s/he has social relationships or belonging to the same community.

Caching Methods

Car manufacturers tend to produce vehicles equipped with OBUs with no memory constraints (i.e., OBUs with large capacity). However, the arrival on the market of these new technologies may take some time. Consequently, nowadays, vehicle users are still using smartphones (with limited buffer size) since there are few cars that are equipped with OBUs. Therefore, given the large volume of data being generated in everyday life, the provision for a content dissemination scheme for soliciting and caching the appropriate set of contents is an effective way to enhance data forwarding in VSNs.

Driver Safety

The number of traffic accidents and traffic violations is increasing due to the use of in-vehicle devices. Therefore, it is necessary to provide a mechanism that minimizes the interaction of the driver with the onboard devices in order to manually share information. The work in [15] has proposed an architecture that enables the vehicle to automatically share information detected autonomously by the sensors in the vehicles.

Driver or Vehicle Behavior

Vehicles are classified, according to their mobility behavior, into three important types:

- *Taxis*: They have random trajectories with low probability to of similar behavior in everyday life.
- *Normal cars*: On weekdays, a normal car usually repeats its paths in the same period to the same destination such as school, work, and so on. On the contrary, during weekends, other destinations, usually chosen for entertainment, are frequently visited.

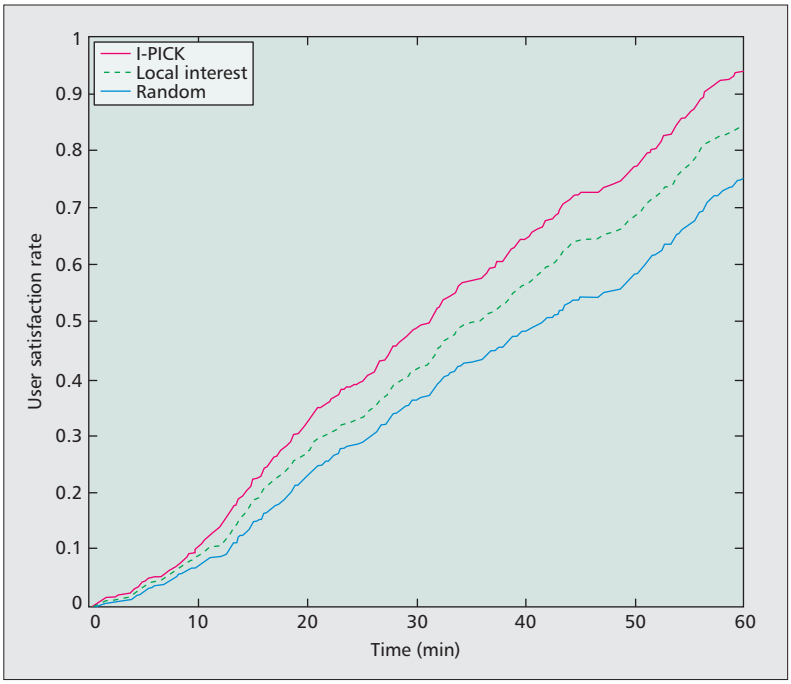


Figure 5. Cumulative satisfaction over time.

- *Buses*: They have a behavior easy to predict since they have a fixed route and schedule. Studies on dissemination algorithms are usually based on the same vehicle type such as buses. Therefore, it is necessary to investigate a general dissemination protocol for VSNs compatible with all types of vehicles.

Conclusion

VSN is emerging as a new hot topic of research in the academic and industrial communities. In VSNs, the concept of social relationships among vehicles can be used to improve the efficiency and effectiveness of content dissemination to meet the requirements of applications and services. This article presents recent achievements in VSNs. It contributes by presenting a taxonomy of existing content dissemination approaches, assisting researchers to advance in the state of the art. In VSNs, a large number of users participate and cooperate to share and access different data. Under this context, this article also presents a framework showing that a user satisfaction feature could be explored to build more efficient content dissemination protocols. Finally, the article discusses open issues and future research directions.

References

[1] S. Smaldone *et al.*, "RoadSpeak: Enabling Voice Chat on Roadways Using Vehicular Social Networks," *1st Wksp. Social Network Sys.*, 2008, pp. 43–48.
[2] F. D. Cunha *et al.*, "Is it Possible to Find Social Properties in Vehicular Networks?," *19th IEEE Symp. Computers and Commun.*, 2014, pp. 1–6.
[3] P. Ranjan and K. K. Ahirwar, "Comparative Study of VANET and MANET Routing Protocols," *Proc. Int'l. Conf. Advanced Computing and Commun. Technologies*, 2011, pp. 978–81.
[4] M. Gerla, C. Wu, G. Pau, and X. Zhu, "Content Distribution in VANETs," *Vehicular Commun.*, vol. 1, no. 1, Jan. 2014, pp. 3–12.

[5] "WAZE: Waze Navigation Service," https://www.waze.com/, last access: Nov. 2014.
[6] W. Sha *et al.*, "Social Vehicle Navigation: Integrating Shared Driving Experience into Vehicle Navigation," *14th Int'l. Wksp. Mobile Computing Sys. and Applications,* 2013, p. 16.
[7] L. Han *et al.*, "Ad-Hoc Voice-Based Group Communication," *Proc. IEEE Int'l. Conf. Pervasive Computing and Communications*, 2010, pp. 190–98.
[8] H. Zhu *et al.*, "ZOOM: Scaling the Mobility for Fast Opportunistic Forwarding in Vehicular Networks," *Proc. IEEE INFOCOM*, 2013, pp. 2832–40.
[9] J. Qin *et al.*, "POST: Exploiting Dynamic Sociality for Mobile Advertising in Vehicular Networks," *Proc. IEEE INFOCOM*, 2014, pp. 1761–69.
[10] R. Fei, K. Yang, and X. Cheng, "A Cooperative Social and Vehicular Network and its Dynamic Bandwidth Allocation Algorithm," *INFOCOM Wksps.*, 2011, pp. 63–67.
[11] N. Cenerario, T. Delot, and S. Ilarri, "A Content-Based Dissemination Protocol for VANETs : Exploiting the Encounter Probability," *IEEE Trans. Intelligent Transportation Systems*, vol. 12, no. 3, 2011, pp. 771–82.
[12] Y. Zhang, J. Zhao, and G. Cao, "Roadcast: A Popularity Aware Content Sharing Scheme in VANETs," *Proc. IEEE ICDCS*, 2009, pp. 223–30.
[13] R. S. Schwartz *et al.*, "On the Applicability of Fair and Adaptive Data Dissemination in Traffic Information Systems," *Ad Hoc Networks,* vol. 13, Feb. 2014, pp. 428–43.
[14] D. Popovici *et al.*, "A Framework for Mobile and Context-Aware Applications Applied to Vehicular Social Networks," *Social Network Analysis and Mining*, vol. 3, no. 3, 2013, pp. 329–40.
[15] I. Lequerica, M. G. Longaron, and P. M. Ruiz, "Drive and Share: Efficient Provisioning of Social Networks in Vehicular Scenarios," *IEEE Commun. Mag.*, 2010, pp. 90–97.

Biographies

Farouk Mezghani (Farouk.Mezghani@enseeiht.fr) received his Engineer degree in telecommunication from the Higher School of Communication of Tunis (SUP'COM), Tunisia, in June 2012. He is currently a Ph.D. student at the National Polytechnic Institute of Toulouse (University of Toulouse, INP-ENSEEIHT) within the Network and Telecommunication Group of the IRIT Laboratory. His research interests include content dissemination, opportunistic networks, and vehicular networks.

Riadh Dhaou (Riadh.Dhaou@enseeiht.fr) is an associate professor at the Telecommunication and Network Division at the ENSEEIHTand a member of IRIT Lab. He received his engineer degree in computer science from the Ecole Nationale des Sciences de l'Informatique (ENSI), University of Tunis II in 1997, and his M.Sc. degree in computer systems from the University of Paris VI in 1998. In November 2002 he was awarded his Ph.D. degree in computer systems, telecommunication and electronics from the University of Paris VI. From 1998 to 2002 he worked as research assistant within the Performance Team of the SAMOVAR Lab at the Institut National des Télécommunications. His research interests include performance evaluation of mobile and vehicular networks, satellite networks, wireless sensor networks, and cross-layer systems.

Michele Nogueira (michele@inf.ufpr.br) is a professor in the Department of Computer Science at the Federal University of Paraná, where she has been since 2010. She received her doctorate in computer science from the Université Pierre et Marie Curie-Sorbonne Universités, Laboratoire d'Informatique de Paris VI (LIP6), in 2009. She was a visiting professor at the ENSEEIHT, IRIT Lab, and a visiting researcher at Georgia Institute of Technology in 2013 and 2009. Her research interests include wireless networks, security, and dependability. She has worked on providing resilience to self-organized and wireless networks by adaptive and opportunistic technologies such as cognitive radio. She was one of the pioneers in addressing survivability issues in self-organized wireless networks, the article "A Survey of Survivability in Mobile Ad Hoc Networks" being one of her prominent scientific contributions. She was a recipient of academic scholarships from the Brazilian Government during her undergraduate and graduate years, and of international grants such as from the ACM SIGCOMM Geodiversity program. She is also an Associate Technical Editor for *IEEE Communications Magazine* and the *Journal of Network and Systems Management*.

André-Luc Beylot (Andre-Luc.Beylot@enseeiht.fr) received his Engineer degree from the Institut d'Informatique d'Entreprise, Evry, France, in 1989 and his Ph.D. degree in computer science from the University of Paris VI in 1993. In January 2000, he received his Habilitation á Diriger des Recherches from the University of Versailles, France. From 1993 to 1995, he worked as a research engineer at the Institut National des Télécommunications, Evry, and from 1995 to 1996 at C.N.E.T. (France Telecom Research and Development), Rennes. From September 1996 to August 2000, he was an assistant professor at the University of Versailles. Since September 2000, he has been a professor at INPT/ENSEEIHT and is member of the IRT Team of the IRIT Laboratory. His research interests are performance evaluation of communication networks, especially with regard to mobile, satellite, and sensor networks.

A Trajectory-Based Recruitment Strategy of Social Sensors for Participatory Sensing

Fei Hao, Mingjie Jiao, Geyong Min, and Laurence T. Yang

ABSTRACT

Participatory sensing, a promising sensing paradigm, enables people to collect and share sensor data on phenomena of interest using mobile devices across many applications, such as smart transportation and air quality monitoring. This article presents a framework of participatory sensing and then focuses on a key technical challenge: developing a trajectory-based recruitment strategy of social sensors in order to enable service providers to identify well suited participants for data sensing based on temporal availability, trust, and energy. To devise a basic recruitment strategy, the Dynamic Tensor Analysis algorithm is initially adopted to learn the time-series tensor of trajectory so that the users' trajectory can be predicted. To guarantee reliable sensing data collection and communication, the trust and energy factors are taken into account jointly in our multi-objective recruitment strategy. In particular, friend-like social sensors are also defined to deal with an emergency during participatory sensing. An illustrative example and experiment are conducted on a university campus to evaluate and demonstrate the feasibility and extensibility of the proposed recruitment strategy.

INTRODUCTION

The popularity of mobile devices and the rapid development of wireless sensing technology advance the emerging of a novel pervasive data sensing paradigm, participatory sensing (PS) [1, 2], which allows citizens to sense their surrounding environment voluntarily with their available sensoring devices (e.g., smart phones) and share the information with other citizens through the existing Internet communication infrastructure. Participatory sensing systems (PSSs) have tremendous potential in various applications, such as environmental monitoring [3], intelligent transportation [4], and route planning [5], because they collect sensing data by virtue of the participatory power of ordinary citizens.

The major difference between PS and traditional sensing lies in each participant being regarded as a sensor, called a social sensor, sensing the surrounding environment to upload data. The analysis ability and mobility of participants in PSSs would greatly reduce the burden on the system and enlarge the geographical coverage of sensing. However, participants, as the data collection carriers of the system, are demanded to sense any time and anywhere, which impedes the wide use of PS. Furthermore, the participants are mostly interested in or related to the sensing campaign. The number of participants is not considerably large and just allows participatory sensing to be applied in a small range.

Consider a real scenario of particulate matter 2.5 (PM 2.5) real-time monitoring in Beijing. In reality, there are insufficient air quality measurement stations in a city due to the expensive cost of building and maintaining such stations. For example, 35 air quality measurement stations are currently established in Beijing. Since these stations are stationary base stations with the traditional network coverage mechanism, they cost lots of money and manpower. Generally, an air quality measurement station needs a certain size of land, a huge amount of money (about US$200,000 for construction and US$30,000 per year for maintenance [6]), human resources to regularly take care of it, and 24-hour-a-day power consumption. These factors greatly limit the number of measurement stations. However, we expect to obtain the measured values of air quality in PSSs through mobile sensing devices held by crowds and further aggregate these values for the purpose of intelligent services supply. In particular, the price of a handheld PM 2.5 sensing device powered by lithium battery (10 W) is US$500. In the worst case, employers (e.g., an environmental protection agency) buy devices for users who are willing to sense the air quality voluntarily/incentively. Roughly 14,210 users can be recruited with the same cost consumed by the traditional sensing system every year for their participatory sensing campaign. From the sustainability point of view, the PS paradigm is better than the traditional sensing paradigm in terms of both cost and energy.

There is some prior research work focused on reputation-based, trust-based, and expertise-based recruitment schemas [1, 8]. In contrast to those existing participant recruitment approaches, this article presents a holistic recruitment strategy that

Fei Hao and Mingjie Jiao are with Huazhong University of Science and Technology.

Geyong Min is with the University of Exeter.

Laurence T. Yang is with Huazhong University of Science and Techonogy, and St. Francis Xavier University.

 0163-6804/14/$25.00

High similarity of trajectories among some participants implies that they might have some social relations in some extents, such as roommates, classmates or family. In this article, we call the participants who have the similar trajectories as a group of "Friends-Like Social Sensors".

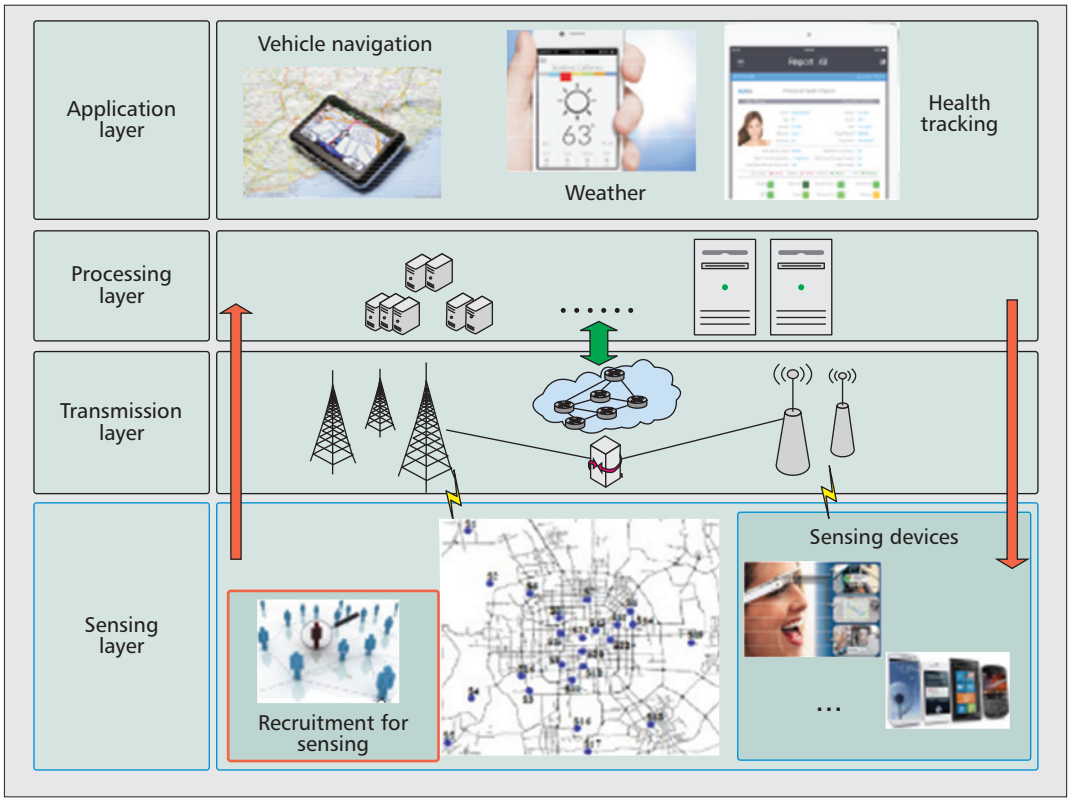


Figure 1. Participatory sensing framework.

considers various impact factors to participatory sensing. Therefore, a PS framework in PSSs is proposed. The proposed framework works as follows: First, the historical trajectories of participants are analyzed for extracting the potential social sensors in the sensing layer; second, the social sensors are dynamically selected in the sensing layer; third, the social sensors voluntarily/incentively sense their surrounding environment and upload these collected data in the servers. Upon this proposed framework, a trajectory-based recruitment strategy of social sensors that considers the availability, trust, and energy of users is devised. To avoid missing sensing data, an emergency selection scheme is also proposed to enhance the usefulness of our recruitment strategy of social sensors. The remainder of this article is structured as follows. We present a PS framework and provide the problem addressed by PS. A trajectory-based recruitment strategy of social sensors for PSSs is proposed. Following an illustrative example, we conclude this article.

Participatory Sensing Framework and Problem Statement

This section provides a typical framework of PS and then presents the problem addressed in this article.

A Framework of Participatory Sensing

Figure 1 presents a typical PS framework in PSSs. This framework is divided into four layers in which different functionalities are enabled. We elaborate the functions and responsibilities of each layer by a bottom-up view approach.

1) Sensing layer: Considering the limited budget and dynamic behaviors of users, a dynamic recruitment strategy of social sensors needs to be proposed in this layer in terms of users' availability, trust value, remaining power of their mobile phones, and emergency context.

2) Data transmission layer: This layer transmits the obtained sensing data to the data center for further processing by the Internet. Before accessing the Internet, mobile users may send the sensing data through either WLAN or a cellular network.

3) Data processing layer: This layer manages the responsibilities of data aggregation, redundant data filtering, data mining, and so on. By processing the obtained sensing data, some relevant services can be provided for the application layer; also, the recruitment strategy might be adjusted according to these data in the sensing layer.

4) Application layer: The various data services in this layer are obtained from the processing layer. For example, the existing services of *vehicle navigation system*, *weather information*, and *health tracking* are widely used in our daily life.

Problem Statement

In this section, the related definitions in PSSs are introduced first. Then the problem statement is described.

Social sensor: In a PSS, a social sensor is actually a participant who is willing to collect data about a particular phenomenon. Therefore, we use participant, user, and social sensor interchangeably in this article.

Trajectory similarity: In daily life, the behavioral trajectories collected by each participant more or less overlap each other. In other words, there is a situation in which some participants appear in the same data collection point (DC-point) at the same time. To quantify this case, trajectory similarity is defined. High similarity of trajectories among some participants implies that they might have some social relations in some cases, such as roommates, classmates, or family. In this article, we call the participants who have the similar trajectories as a group of *friend-like social sensors*.

Participatory sensing system: The essence of participatory sensing is data collection and interpretation. Participation requirements allow a campaign organizer (service provider) to recruit participants who have a certain level of experience or are available in a certain time-spatial space. Participation metrics include:

- The number of campaigns volunteered for
- The number of campaigns accepted
- The number of campaigns participated in
- The number of campaigns abandoned

Individual metrics can be associated with other information about a campaign, such as size, lifetime, and type of sensing required; for example, some potential participants who have been selected for traffic sensing campaigns in the past three months in a certain area.

Generally, a successful PS campaign is dependent on two main issues:

- How to build an efficient recruitment strategy of social sensors
- How to make an incentive mechanism for motivating these social sensors to participate in the sensing campaign based on the recruitment strategy

Our work in this article focuses on devising a recruitment strategy in which each social sensor is dynamically selected and assigned to a set of DC-points where data should be collected. In this section, we formally describe this problem as follows.

(**Problem Statement**) Given an area and a group of possible social sensors U with their mobile traces, the entire recruitment strategy of social sensors is composed of the following technical aspect: for a campaign $C(G, T)$, with G as the set of the DC-points, and T as the set of time of data sensing and collection determined by the campaign requirements. Thus, the recruitment problem in the recruitment layer is to dynamically select each participant $u \in U$ to any DC-point located in G, such that the location of u is closer to the DC-point in its sensing range g than to that of any others in U. Then those recruited participants will carry out $C(G, T)$ at the required time and location for campaign organizers.

A Trajectory-Based Recruitment Strategy for Social Sensors

Participation sensing can effectively replace stationary base stations by recruiting participants. By predicting the trajectory data of participants, it can help us to select the appropriate participants to join the PS campaign.

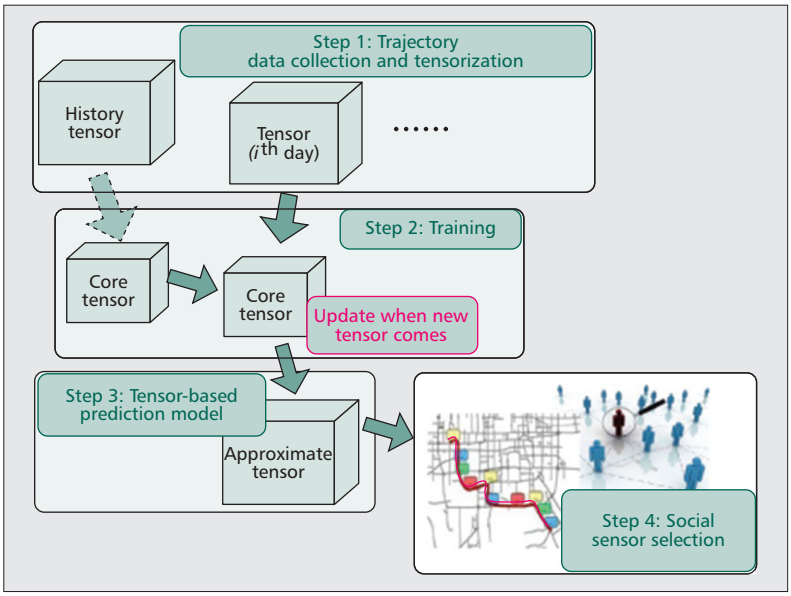


Figure 2. An overview of our recruitment strategy of social sensors.

Big Picture

The proposed recruitment strategy of social sensors works within a given monitoring area with M DC-points and N users. As shown in Fig. 2, our social sensor recruitment strategy contains the following three steps:

***Step 1*: Trajectory data collection and tensorization**: The trajectory data of participants associated with users, time, and location is represented with a tensor $\chi \in \Re^{I_t \times I_g \times I_u}$. We collect the trajectory data within i days. The trajectory data in the ith day is a tensor χ_i. Therefore, the collected data is represented by a time-series tensor $\chi^T = \{\chi_1, \chi_2, \ldots, \chi_i\}$.

***Step 2*: Tensor-based data training**: The time-series tensor χ^T is trained by the dynamic tensor analysis (DTA) approach [7]. Then an approximate tensor $\tilde{\chi}$ is obtained.

***Step 3:* Prediction and friend-like social sensors identification**: Based on the obtained approximate tensor, the user's future moving patterns or expected arrival locations are predicted. Furthermore, Euclidean distance is adopted to measure the similarity among the moving patterns of social sensors within a period. Finally, we cluster those users who have high similarity of moving patterns into a group of friend-like social sensors.

***Step 4:* Social sensor selection**: Generally, we dynamically select the optimal social sensors who can satisfy the availability, trust, and energy constraints at each time. According to the historical trajectory, we infer the availability of each social sensor appearing near the DC-point using the approximate tensor. During the PS interaction, the trust and energy of each social sensor are taken into account timely for adjusting the selection results. In particular, when an emergency happens, such as sensing devices with lower power, urgent personal affairs, and so forth, the participants contained in the group of friend-like social sensors will become the selection candidates. Then the selected social sensors are stimulated to participate in a given sensing campaign.

These dimensions constructed as a tensor are important for representation, processing, and storage of PSSs. Hence, this strategy can help us to discover the potential semantic relationships from those data and provide intelligent services for the participatory sensing campaign.

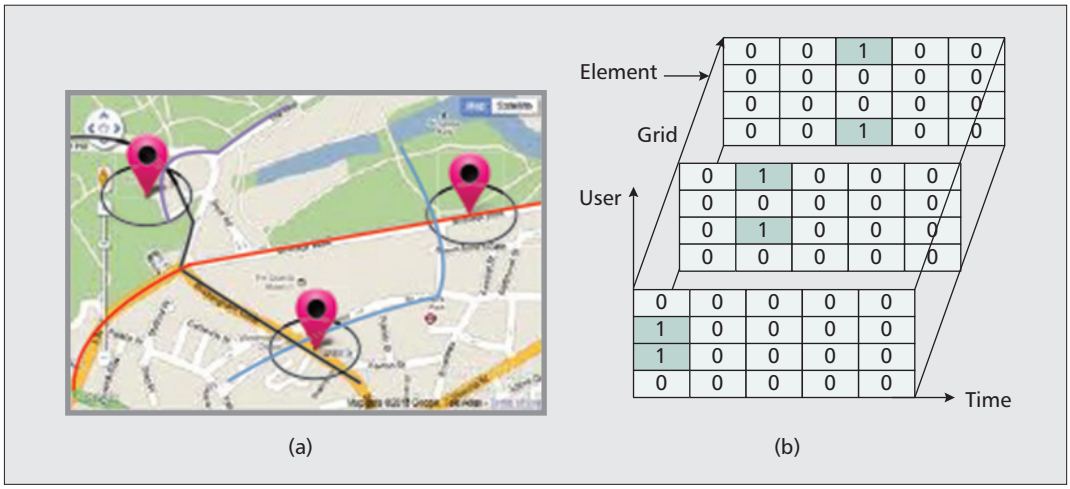


Figure 3. Trajectory data tensorization: a) monitoring area with grids; b) tensor representation for trajectory.

With this overview of our proposed recruitment strategy of participants, the following sections present the detailed strategy with the tensor-based Dynamic Tensor Analysis (DTA) algorithm. A tensor, as a type of high-dimension matrix that governs the correlations among these dimensions, is widely used in many applications. In PSSs, the trajectory data of a certain period is regarded as a type of high-dimensional tensor that is associated with users, time, and location. These dimensions constructed as a tensor are important for representation, processing, and storage of PSSs. Hence, this strategy can help us discover the potential semantic relationships from those data and provide intelligent services for the PS campaign.

Trajectory Data Tensorization and Collection

For a given monitoring area, there are N users who are going about their daily activities in this area. In order to tensorize the trajectory data of these N users, we first position and determine the virtual sensing range of those M DC-points, and then collect the daily trajectory data with k time intervals. Apparently, the trajectory data generating from PSSs is mainly composed of three dimensions: *users*, *time*, and *locations*. (Note that the longitude and altitude of a certain location correspond to a certain pre-partitioned grid). The element of the daily trajectory data can be described as a 4-tuple $a = <T, G, U, V>$, where T is time, G refers to the grid in which the users are staying, U denotes a certain user, and V is the element's value. This 4-tuple corresponds to a 3-order tensor as $\chi \in \Re^{I_t \times I_g \times I_u}$, where $\Re$ is defined on the real number domain. I_t, I_g, and I_u refer to time, location grids, and users. $I_t \times I_g \times I_u$ denotes the Cartesisan product of each individual domain. The value of each element $x(t_k, g_m, u_n)$ in the 3-order tensor represents the likeliness of user u_n staying in grid g_m at time t_k, which is obtained by GPS-enabled devices.

In other words, a user can only belong to a certain time. Hence, the constructed tensor including the daily trajectory data is very sparse, as shown in Fig. 3.

Before the trajectory data training, we construct the timeseries tensor $\chi^T = \{\chi_1, \chi_2, \ldots, \chi_i\}$ by collecting i days' daily trajectory data of users.

Trajectory Data Training

Dynamic tensor analysis [7] is an efficient algorithm for dynamically revealing the hidden correlations among the dimensions (e.g., time, users, and locations in PSSs) of the tensor. Therefore, we adopt the DTA algorithm to analyze the time-series tensor X^T and mine the potential trajectory patterns of users in PSSs.

An initial tensor can be matricized in several modes, which are determined by their orders. For example, tensor $\chi \in \Re^{I_t \times I_g \times I_u}$ has three unfolding matrices that can be decomposed into a projection matrix $U_{(d)}$ and an energy matrix $S_{(d)}$ via singular value decomposition (SVD) for the corresponding mode d. Then a covariance matrix $C_{(d)}$ can be calculated with $U_{(d)}$ and $S_{(d)}$.

The DTA algorithm processes each mode of the tensor continuously. Importantly, the $C_{(d)}$ is updated as $C_{(d)} \leftarrow \lambda C_{(d)} + X_{(d)} X_{(d)}^T$, where $\lambda \in [0, 1]$ is a forgetting factor regarded as the predictable information aggregator of time series data. In other words, recent timestamps are more important than those far in the past. Then we decompose the above updated covariance matrix $C_{(d)}$ to obtain the principal eigenvectors that are used for computation of the core tensor in the following step. The aforementioned method for calculating the core tensor is just a training process focused on one timestamp. Clearly, the core tensor is updating dynamically. The calculation of the core tensor is equivalent to learning the historical tensors.

Eventually, an approximate tensor $\tilde{\chi}$ used in the tensor-based prediction model is the product of the core tensor and three metrics.

Trajectory Prediction and Friend-Like Social Sensor Identification

The approximate tensor $\tilde{\chi} \in \Re^{I_t \times I_g \times I_u}$ is actually an information aggregator of the results learned from the previous time-series tensors. Each element value in $\tilde{\chi}$ denotes the likeliness of a user

appearing in a certain grid at a certain time. Therefore, the future trajectory can be predicted by the approximate tensor.

Furthermore, we attempt to identify the friend-like social sensors by using Euclidean distance, which powerfully measures the distance between two corresponding points located in the trajectories of the 3-order tensor. In other words, the similarity between any two trajectories of different social sensors can easily be estimated according to the following Euclidean distances:

$$sim\left(\widetilde{\chi}(:,:,u)\widetilde{\chi}(:,:,v)\right) = \frac{\sum_{i=0}^{n}\frac{1}{Dis\left(\widetilde{\chi}(g_i,:,u)\widetilde{\chi}(g_i,:,v)\right)+1}}{n}$$

where the $\widetilde{\chi}(:, :, u)$ denotes the trajectory of social sensor u in the future; $Dis(X, Y)$ refers to the Euclidean distance between vector X and Y; and $sim(\widetilde{\chi}(:, :, u); \widetilde{\chi}(:, :, v)) \in (0, 1]$.

Based on Eq. 1, if the similarity between two trajectories of social sensor is greater than a given threshold γ, that is, $sim(\widetilde{\chi}(:, :, u); \widetilde{\chi}(:, :, v)) \geq \gamma$, u and v are regarded as the friendlike social sensors reciprocally.

Social Sensor Selection

Since some users may be selected for the settled time and grid, we have a selection strategy to pick up better volunteers. At each time, we can dynamically select well suited social sensors who can satisfy the constraints of availability, trust, and energy. For example, the availability can be reasoned by the approximate tensor. We rank the likeliness value of the users on the decrease, and choose the top k users who can be the potential social sensors. Based on this idea, the following two strategies:

- *Basic selection strategy*
- *Multi-objective selection strategy*

are devised, respectively.

Basic Selection Strategy — As a basic selection strategy, availability is a critical factor to be considered. For the targeted grid at a given time, the availability-based selection strategy of social sensors is dependent on the likeliness ($x'(t, g, u) \in \widetilde{X}$) rank of the users appearing in the targeted area at that time.

Multi-Objective Selection Strategy — From the data reliability point of view, the factors of energy and trust used to evaluate the reliability of participatory sensing between participants and server [8] are jointly taken into account; then a multi-objective selection strategy is devised and formalized as follows:

$$\begin{aligned} &\max\ \alpha x'(t, g, u) + \beta T(u) + \gamma E(u) \\ &s.t. \quad T(u) \geq \theta_{trust} \\ &\qquad E(u) \geq \theta_{energy} \end{aligned}$$

On one hand, Fig. 5a illustrates that if a social sensor candidate in the top k social sensor list has already participated and been selected in the sensing campaign more frequently, it implies that sensor u has a higher trust value, denoted as $T(u)$. Then we extract the social sensors whose trust values are greater than a given threshold θ_{trust} which is determined by service providers. On the other hand, service providers hope that the possible social sensors u have enough energy remaining in their mobile devices, $E(u)$. The energy consumption for each participant during the sensing period follows an exponential decay trend [10] $E(t) = E(0)e^{-2t}$ as shown in Fig. 5b, where $E(t)$ indicates the residual energy at time t, $E(0)$ is an initial energy in their mobile devices. Thus, each social sensor should report back the remaining energy information and sensing data. Similarly, another threshold θ_{energy} is adopted for further refining in order to guarantee continuous and reliable data sensing and communication. Note that the weighted parameters α, β, and γ can be learned with the least square method [9]. In particular, the above multi-objective selection strategy is degraded to the basic selection strategy of social sensors.

In the real world, an emergency inevitably occurs during PS, such as insufficient power of sensing devices or urgent affairs of u; then we can select the friend-like social sensors of u to ensure that the PS campaign continues.

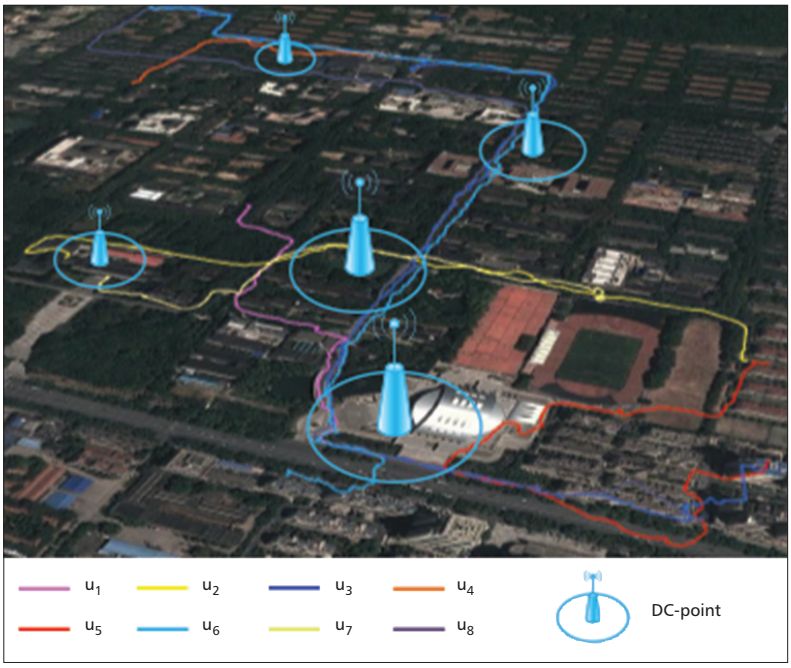


Figure 4. Trajectory of five typical volunteers and three assistant volunteers on 5 November.

An Illustrative Example

In this section, we present an illustrative example of participatory sensing on the HUST campus to evaluate the feasibility and effectiveness of the basic recruitment strategy.

Setup

For a given monitoring area, we first position and determined the virtual sensing range of five DC-points. We first created a public microblog ID and regard it as a sensing data monitoring platform. We collected the GPS location and

Since some users may be selected for the settled time and grid, we have a selection strategy to pick up the better volunteers. At each time, we can dynamically select the well-suited social sensors who can satisfy the constraints of availability, trust and energy.

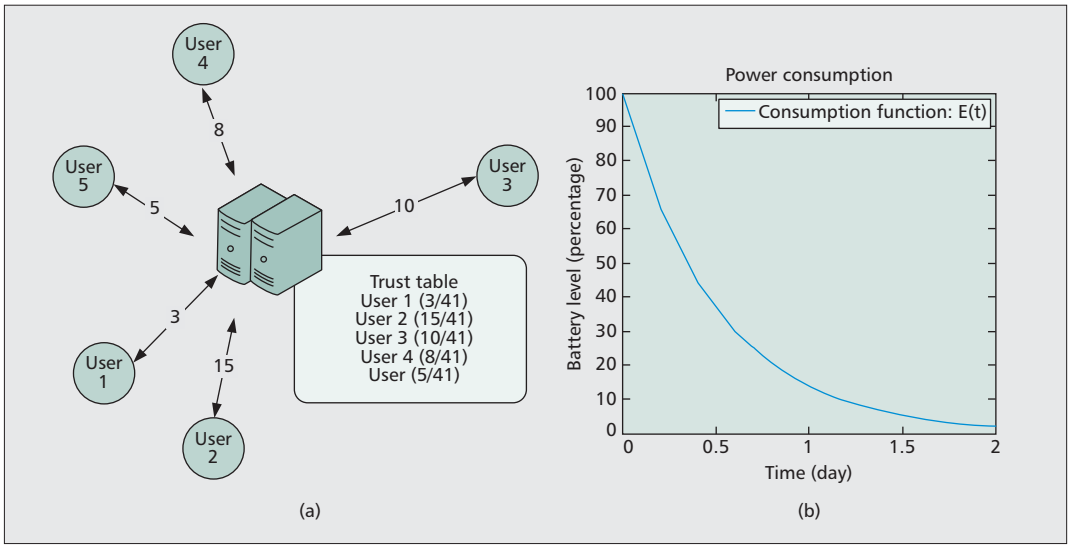


Figure 5. Evaluation of trust and energy consumption: a) trust evaluation; b) energy decay curve.

noise information every 2 min from 8:00 to 8:40 a.m. from a number of volunteers at our university from September to November 2013 according to their social media feedback (texts, images) interacted with our public microblog ID. Thus, these collected daily trajectory data can be constructed as a 3-order tensor that includes users, grids, and time dimensions. The detailed steps can be found at our website.[1]

Results and Discussions

After data collection, the time-series tensor χ^T of trajectory data is trained by DTA algorithm. For illustration and visualization purpose, we choose the trajectory data gathered by 5 typical volunteers (u_1, …, u_5) between 2 and 8 November as shown in Fig. 4. In our experiments, these five participants are the representative social sensors in five groups of friend-like social sensors. Table 1 shows the expected participants to be recruited at four different times in five grids including pre-deployed DC-points. Clearly, the campaign organizer will recruit users u_1, u_4 as the potential social sensors in grid g_1 at time t_7. The likeliness value reflects the possibility of users who are to be recruited. However, user u_5 will not be considered in this participatory campaign due to his low likeliness. For example, user u_4 should first be considered as a social sensor in grid g_1 at time t_6 because of the higher likeliness value compared to that of u_1.

As mentioned above, trust is one of the considerations in multi-objective selection strategy and an important personalized factor [8]. Obviously, u_1 is more likely available in grid g_{13}, which includes pre-deployed DC-points from time t_9 to t_{11}. If we choose u_1 as the social sensor at that period, more interactions are benefit to enhancing the trust value that may assist the further selection next time. Since diverse sensing devices (e.g., PM2.5 sensing devices, smartphones) are held by our volunteers, estimating energy consumption of these devices is becoming a challenge. We will study the multiobjective selection strategy with energy consideration in future work.

Campaign C <*grid*, *time*>	Participants user (likeliness)
C <g_1, t_6> C <g_1, t_7>	u_4 (0.963), u_1 (0.152), u_3 (0.017) u_4 (1.036), u_1 (0.901), u_5 (0.036)
C <g_2, t_5>	u_1 (0.977), u_2 (0.001)
C <g_3, t_{10}> C <g_3, t_{11}>	u_1 (1.052), u_3 (–0.102) u_1 (0.938)
C <g_4, t_{10}> C <g_4, t_{11}> C <g_4, t_{14}>	u_5 (1.131), u_2 (–0.230) u_5 (0.958) u_5 (0.230), u_1 (0.210)
C <g_5, t_{15}>	u_5 (1.118)

Table 1. Expected participants to be recruited in various PS campaigns.

In reality, PSSs usually suffer several disruptions due to the lower power of sensing devices and personal urgent affairs. Hence, an emergency selection strategy of social sensors needs to be devised for achieving a reliable and accurate participatory sensing campaign. Our proposed emergency selection strategy is to choose the participants contained in the group of friend-like social sensors in the event of emergency. To evaluate the feasibility of the emergency selection strategy, users u_6, u_7, and u_8 are taken as the assistant social sensors who might be selected to be the friend-like social sensors. In Fig. 4, if user u_3 suddenly terminates her sensing temporarily due to some urgent personal affairs, the PSS should receive this feedback and select the potential social sensors u_6 and u_8 from the group of friend-like social sensors of u_3. Actually, these results are reasonable because users u_6, u_8, and u_3 have the relationships of both *roommates* and *classmates*.

[1] http://epic.hust.edu.cn/ps

Conclusions

To realize a novel sustainable sensing, this article investigates the social sensor recruitment problem in participatory sensing systems. We first present a framework of PSSs. Since the recruitment layer in the proposed framework has not been investigated yet, this article focuses on the issue of recruitment of social sensors and proposes a trajectory-based recruitment strategy of social sensors for participatory sensing. Specifically, the collected trajectory data of users within a period are constructed as a 3-order time-series tensor. Furthermore, the DTA algorithm is adopted to learn this time-series tensor and predict the future trajectory information that supports the basic selection strategy based on availability. Finally, the proposed selection strategy is evaluated with an illustrative example conducted on the HUST campus. The proposed framework and associated techniques pave a way for achieving intelligent services in PSSs by the virtue of the participatory power of the selected well suited participants.

References

[1] X. O. Wang *et al.*, "ARTSense: Anonymous Reputation and Trust in Participatory Sensing," *Proc. INFOCOM '13*, 2013, pp. 2517–25.

[2] S. Reddy, D. Estrin, and M. Srivastava, "Recruitment Framework for Participatory Sensing Data Collections," *Proc. Pervasive '10*, 2010, pp. 138–55.

[3] V. Kotovirta *et al.*, "Participatory Sensing in Environmental Monitoring–Experiences," *Proc. IMIS '12*, 2012, pp. 155–62.

[4] P. Zhou, Z. Chen, and M. Li, "Smart Traffic Monitoring with Participatory Sensing," *Proc. SenSys '13*, 2013, pp. 26:1–26:2.

[5] S. Reddy *et al.*, "Biketastic: Sensing and Mapping for Better Biking," *Proc. SIGCHI '10*, 2010, pp. 1817–20.

[6] Y. Zheng, F. Liu, and H. P. Hsieh, "U-Air: When Urban Air Quality Inference Meets Big Data," *Proc. KDD '13*, 2013.

[7] J. Sun, D. Tao, and C. Faloutsos, "Beyond Streams and Graphs: Dynamic Tensor Analysis," *Proc. KDD '06*, 2006, pp. 74–383.

[8] H. Amintoosi and S. S. Kanhere, "A Trust-Based Recruitment Framework for Multihop Social Participatory Sensing," *Proc. DCOSS '13*, 2013, pp. 266–73.

[9] I. Yeo, C. C. Liu, and E. J. Kim, "Predictive Dynamic Thermal Management for Multicore Systems," *Proc. ACM/IEEE Design Automation Conf.*, 2008, pp. 734–39.

[10] A. P. Miettinen and J. K. Nurminen, "Energy Efficiency of Mobile Clients in Cloud Computing," *Proc. 2nd USENIX Conf. Hot Topics in Cloud Computing*, 2010.

Biographies

Fei Hao (fechao@gmail.com) is an assistant professor at Huazhong University of Science and Technology. He received his B.S. and M.Sc. degrees in the School of Mathematics and Computer Engineering from Xihua University, Chengdu, China, in 2005 and 2008, respectively. He was a research assistant at Korea Advanced Institute of Science and Technology and Hangul Engineering Research Center. He is studying toward his Ph.D. degree in the Department of Computing at the University of Bradford, United Kingdom. He has published over 30 research in international and national journals as well as conferences. His research interests include social computing, big data analysis and processing, and mobile cloud computing.

Mingjie Jiao (mingjie.v@gmail.com) is an M.Sc. student in computer science and technology at Huazhong University of Science and Technology. Currently, he is working at Baidu on personal trajectory data mining.

Geyong Min (g.min@exeter.ac.uk) is a professor of high performance computing and networking in the Department of Mathematics and Computer Science at the University of Exeter, United Kingdom. He received his Ph.D. degree in computing science from the University of Glasgow, United Kingdom, in 2003, and his B.Sc. degree in computer science from Huazhong Univerity of Science and Technology, China, in 1995. His research interests include future Internet, computer networks, wireless communications, multimedia systems, information security, high-performance computing, ubiquitous computing, modeling, and performance engineering.

Laurence T. Yang (ltyang@ieee.org) received his B.E. degree in computer science and technology from Tsinghua University, China, and his Ph.D. degree in computer science from the University of Victoria, Canada. He is affiliated with School of Computer Science and Technology, Huazhong University of Science and University, China, as well as with Department of Computer Science, St. Francis Xavier University, Canada. His research interests include parallel and distributed computing, embedded and ubiquitous/pervasive computing, and big data. He has published more than 200 papers in various refereed journals (about 40 percent in IEEE/ACM transactions and journals, and the others mostly in Elsevier, Springer, and Wiley journals). His research has been supported by the National Sciences and Engineering Research Council of Canada, and the Canada Foundation for Innovation.

The proposed selection strategy is evaluated with an illustrative example conducted at HUST campus. The proposed framework and associated techniques pave a way for achieving intelligent services in PSSs by the virtue of the participatory power of the selected well-suited participants.

Security and Performance Challenges for User-Centric Wireless Networking

Pantelis A. Frangoudis and George C. Polyzos

ABSTRACT

User-centrism has emerged as a disruptive new communication paradigm. In this article, we lay out its basic principles, study the key factors that have given rise to its adoption, and focus on the new set of challenges it brings about in various aspects of *wireless* networking. We study user-centric solutions on a case-by-case basis, along the dimensions of wireless access, provision of communication services, and wireless network management. We tackle specific security and performance challenges by designing and implementing architectures for secure VoIP communication tailored to user-centric wireless networks, and for robust user-driven wireless topology monitoring, a critical network management task. In both cases, we quantify the trade-off between security and performance, showcasing the potential of relying on users to carry out traditionally provider-centric tasks.

INTRODUCTION

From the lower communication layers up to the content layer, where user-generated content vastly increases in volume and popularity, it is becoming evident that the traditional "user as a consumer" view of networking is evolving to a disruptive new paradigm where the user emerges with an empowered role. The user-centric shift is especially pronounced in the area of wireless networking, where it finds expression in multiple dimensions.

This evolution is supported by various technology developments. User equipment is becoming more powerful, and comes with an abundance of storage, sensing, and multimedia generation and playout capabilities. At the same time, wireless technologies operating in unlicensed spectrum, such as Wi-Fi, have become ubiquitous. These have the potential of bringing up new disruptive communication and service architectures. More than ever, though, a user-centric shift is also necessitated by the challenges brought by the vast increase in mobile data traffic [1]. This calls for alternative content distribution schemes, exploiting user capabilities to shift the load off the cellular network and the backhaul of mobile operators.

In this article, we study the developments that have rendered the user a critical stakeholder in the wireless networking ecosystem, defining a basic set of principles of user-centrism that lay the foundations for reconsidering the role of users in three different wireless networking aspects: wireless access provision, communication services, and network management. Our particular focus is to expose some new fundamental challenges and trade-offs that user-centric solutions face in a case-by-case manner.

To this end, we first present purely user-centric solutions for real-time communication tailored to user-centric wireless access networks, tackling specific security challenges and experimentally demonstrating their performance bounds.

Then we focus on the role of users in wireless network monitoring, a critical task for optimized wireless network configuration and management. In this case, we present an architecture for crowdsourced Wi-Fi topology discovery, combating attacks by users who submit fraudulent information, and quantifying its advantages over pure infrastructure/provider-centric schemes. Figure 1 presents our overall view toward user-centric wireless networking.

THE EMERGENCE OF USER-CENTRISM IN WIRELESS NETWORKING

KEY FACTORS

A number of key developments and observations are indicative of the shift toward a user-centric view of networking, and form the basis for proposing user-centric approaches to issues that were traditionally tackled in a provider/operator-oriented manner.

Increased User-Based Wireless Coverage — Wi-Fi signals pervade modern urban spaces. Most of these networks are managed by individuals, in addition to those operated within corporate premises, campus environments, and other public places. Increased Wi-Fi coverage gives rise to the question of whether such user-provided infrastructure could be harnessed to offer a low-cost ubiquitous wireless access solution to complement cellular services.

Pantelis A. Frangoudis is with INRIA Rennes-Bretagne Atlantique.

George C. Polyzos is with Athens University of Economics and Business.

0163-6804/14/$25.00 © 2014 IEEE

Flexible Technologies for the Wireless Home Network — Low-cost off-the-shelf wireless equipment for the home network is capable of performing far more tasks than simply forwarding user traffic to/from the Internet. Home wireless routers powered by open source software have both the necessary flexibility to install custom software, and potential spare memory and CPU cycles to run more demanding applications. User equipment with increased power and flexibility, sometimes supported by telecom operators who utilize it for deploying femtocells or extending their coverage via user Wi-Fi networks, is critical for the emergence of users as *prosumers* of wireless connectivity.

Versatile Technologies at the User End — Apart from their increase in processing power and storage capacity, handheld user devices come equipped with multiple network interfaces (Wi-Fi, cellular, Bluetooth), high-quality displays, but also versatile sensing capabilities via motion sensors, cameras, and GPS receivers. These technologies, combined with the inherent user mobility, make handheld devices powerful platforms to acquire and communicate information about user environment and context, thus giving rise to *crowdsourcing* [2].

The Rise of Crowdsourcing — A shift toward exploiting user resources and advanced communication, computation, and sensing capabilities to "outsource" tasks to potentially anonymous/pseudonymous crowds has become apparent. Such tasks, which would traditionally need significant investment in infrastructure and time, range from urban sensing to collecting information about the radio environment that can be used for network optimization purposes.

User-Generated Content — It is evident that user-generated content takes up a significant share of today's Internet traffic. Such traffic includes web-accessible multimedia content (e.g., videos captured or authored by users, photos, etc.) and interactions via social networking media.

Principles

The above key factors can inspire user-centric wireless networking approaches, building on the common ground of a core set of principles.

The User at the Center — Our prime principle is that the problems we address should be viewed under a user-centric perspective and the solutions proposed should promote the role of users, exploiting and showcasing user empowerment.

Open Access and Participation — User-centric solutions should facilitate and promote open and voluntary user participation, for example, users pooling their home WLANs with the common goal of achieving wider network coverage and enjoying low-cost wireless access when nomadic.

To facilitate organic growth and lift entry barriers to become a service provider, joining a user-centric network (UCN) should not involve the complexity of setting up contracts; rather, a loose and decentralized user identification scheme is desirable.

Decentralization and Distribution of Tasks — Even though some functions (e.g., addressing and naming) may be centrally managed, decentralization naturally emerges in UCNs at many layers. In this article, we demonstrate decentralized designs for various traditionally centralized networking tasks.

Security, Trust, and User Rationality — Mechanisms and protocols for UCNs should not assume that users are benevolent. Often, rational behavior that can lead a user to strategic decisions which may violate protocols, or even pure malice, emerge. Such behaviors can lead to attacks the system designer should tackle, and effective mechanisms taking into account the underlying trust relationships between UCN entities should be in place to ensure that potential attackers cannot bring the system to suboptimal operating points.

Low-Cost Operation — This is mandated by the need to amass infrastructure based on private resource contributions of individuals who own and operate inexpensive home (not professional) equipment. Decentralized service architectures can then be laid over this infrastructure and accessed in a peer-to-peer manner. This can enable, for example, free communication services over wireless UCNs. This principle is in sharp contrast with the Internet service provider (ISP)-centric viewpoint; instead of a few large providers, many *micro-operators* could offer a complementary best-effort service at minimal cost.

New Challenges

Building infrastructure and services based on user contributions poses a new set of challenges. First, user-provided equipment is typically inexpensive and resource-constrained, standing at the opposite side from powerful and costly infrastructure deployed by operators. Second, users often do not have the technical expertise, cannot spare the resources and time, and lack central control when it comes to configuration decisions, optimizations, and planning operations. Therefore, user-provided services, be they connectivity-related, application-oriented, or having to do with content and information provision, are assumed to operate in a self-organizing best-effort style, also due to the unpredictability and variability in user behavior and participation. Designing and implementing user-centric protocols, mechanisms, and services thus involves addressing significant *performance* and *reliability* challenges.

Furthermore, crowds of users pooling their resources, executing a distributed task, or offering a service cannot always be assumed to be trustworthy. They are typically not legally bound by contracts and service level agreements, and are expected to behave strategically, without excluding the potential of purely malicious behavior. In some of the use cases we study, user identification is not always assumed to be strong.

A number of key developments and observations are indicative of the shift toward a user-centric view of networking, and form the basis for proposing user-centric approaches to issues that were traditionally tackled in a provider/operator-oriented manner.

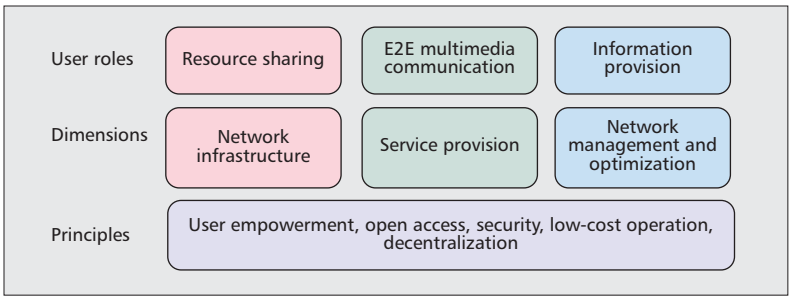


Figure 1. Aspects of user-centric wireless networking.

While this enhances anonymity and privacy, and reduces identity management overhead, it calls for careful design of mechanisms and additional *security* measures.

In the next sections, we address these challenges presenting user-centric solutions in different areas of wireless networking.

From Crowdsourced Wireless Access to Socially Aware, User-Centric Content Delivery

Since the emergence of the IEEE 802.11 family of standards, and as the technology matured, a trend toward open wireless access has been evident. Operation in unlicensed spectrum and the low cost of WLAN equipment, coupled with the enthusiasm and self-organizing spirit of some users, have helped build community-based wireless access schemes for open connectivity [3]. A typical manifestation of such schemes is crowdsourced Wi-Fi access, that is, community-based Internet access over WLANs operated by a crowd of micro-operators, who are typically residential WLAN owners. This paradigm has received both research [4] and commercial attention.[1]

A major issue here is to design appropriate mechanisms to stimulate user participation and cooperation. The proper incentives need to be in place so that a user shares her Internet connection with other community members. Without protection mechanisms for resource contributors, *free riding* will prevail. Users will tend to consume without providing, and cooperation will collapse. We specifically tackled these issues, focusing on achieving distributed trust and accountability in a fully anonymous and decentralized setting [4], contrary to the Fon model, where a centralized authentication, authorization, and accounting scheme is assumed. Our distributed algorithms successfully tackle specific attacks to the accounting mechanisms, attempting to match user consumption with contribution, promote good contributors, and exclude free riders, thus encouraging cooperation and sharing.

There are further wireless and mobile networking issues that can be viewed under a user-centric prism. With the evolution of the Internet toward an infrastructure for massive content delivery and the vast expected increase in mobile multimedia traffic [1], it would be advantageous for network operators and content providers to exploit the communication capabilities and storage capacity of users, making them *active components* of the content delivery process. Mobile data offloading schemes exploiting user storage and device-to-device (D2D) communication are promising solutions in this direction.

Serving user requests from alternative communication channels can reduce the operating and capital expenses of mobile network operators. D2D communication can be at the core of such an approach. Taking advantage of opportunistic direct local connectivity, an approach particularly suitable for delivering delay-tolerant content, efficient delivery structures can be built, benefitting all involved entities: Users can enjoy better quality of experience (QoE), network operators can save on resources on their radio network and backhaul, and content providers can reduce the load on their data centers, thus achieving shorter response times and energy savings. Cooperation scenarios among all stakeholders are thus likely to emerge and could become a topic for extensive study, from both a technical point of view and a socioeconomic one.

User *social context* is critical for the design of such schemes. Previous studies [5] indicate that people who are close in physical and social space tend to have more encounters. This dimension can be exploited to more efficiently plan offloading decisions for delay-tolerant, potentially user-generated, content: Transmissions over the cellular network can be reduced and a user can receive the requested content in the near future over a direct connection with a socially proximate user (e.g., using Wi-Fi or Bluetooth). Such solutions have already begun to be explored [6].

Detecting social context is, however, not trivial. Information from multiple sources can be utilized to this end: application-layer information from existing social networks, but also historical information about past user encounters or communication. Privacy concerns then naturally arise, since users expose information that could be considered sensitive. Privacy-enhancing technologies should be applied, and a trade-off to explore is between the level of user privacy and content delivery performance (and, in turn, user experience).

The incentives of each party to participate in the content delivery process are not always straightforward. For instance, proper rewarding mechanisms may need to be in place to stimulate the participation of users in a cooperative caching scheme where they would contribute device storage space. Especially when it comes to data offloading using D2D communication, energy issues come to play: While it is clearly more energy-efficient to receive a large piece of content over, say, Bluetooth than a (macro-) cellular link, the energy cost for a user to relay content to others should be considered.

User-Centric Secure Communication Services

With user-based wireless access schemes in place, real-time communication services can be set up in a manner compatible with the user-centric spirit. Such service architectures can add value to the infrastructure contributions of

[1] Major telecom operators in many European countries offer the option to their subscribers to roam across one another's home WLANs, sometimes striking partnerships with Fon (http://fon.com), a well-known mediator for community-based Wi-Fi access.

micro-providers. However, due to its particularities, this environment comes with a new set of requirements and constraints.

Requirements

We assume an environment where roaming users, accessing the Internet over community or other public wireless hotspots, wish to set up secure communication. We focus on the deployment of a user-centric VoIP application, but our approach can also be applied to other services (e.g., video). The principles of user-centrism are expressed in the following set of requirements:

- Minimal dependence on centralized infrastructures: Core operations, such as security management, should be controlled by the users, and the role of centralized entities should be limited only to some necessary functionality (e.g., to assist in user discovery and rendezvous).
- Protection from untrusted peers: Nomadic users typically connect to the Internet through untrusted user-provided networks, and their traffic is susceptible to interception if not protected; micro-providers, on the other hand, share their resources with untrusted individuals, who could engage in malicious activities masking behind the micro-provider's network. Furthermore, a level of location privacy is desirable for roaming users.
- Operation on user-provided equipment: This requirement is mandated by the need to operate in an autonomous manner and at low cost, but reveals a significant performance challenge. User equipment is typically resource-constrained, which can cause performance degradation when called to carry out demanding tasks.

Note that we consider only low-mobility nomadic users; we defer mobility management issues to future work.

VPN-Based Solutions

To answer these challenges, the solution comes through tunneling-based schemes: Each user can set up a secure virtual private network (VPN) to a trusted gateway and route his Internet traffic through it. Adopting a user-centric approach, in our prior work [7] we have proposed to utilize a user's home Wi-Fi router as her trusted VPN gateway, and have designed appropriate mechanisms for call setup and tunnel management. Our architecture (Fig. 2) takes advantage of the fact that often off-the-shelf Wi-Fi equipment is Linux-powered, offering the flexibility to install versatile software on it.

This scheme uses centralized infrastructures only for *rendezvous*. For example, a secure VoIP call could be initiated utilizing an external channel, such as the GSM network: A user can send an SMS to a peer notifying him of her home gateway address and other call parameters, and the callee can respond directly with the VoIP stream. Other options, based on dynamic Domain Name Service (DNS) or Session Initiation Protocol (SIP), are also possible [7].

A VPN-based solution offers protection of user traffic from an untrusted access provider, at the same time protecting the micro-provider himself from potential malicious activities originating from the untrusted visitor. Also, communicating endpoints hide their location, since peers direct their traffic to each other's VPN gateways.

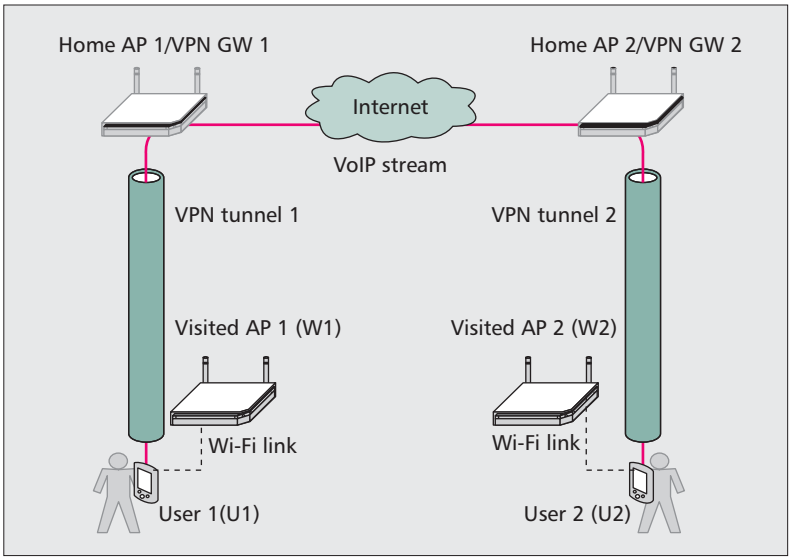


Figure 2. A user-centric secure VoIP service for roaming users.

Others have also addressed these challenges with tunneling-based designs. Sastry *et al.* [8] propose that users form a cooperative acting as a pool of VPN gateways, and roaming community members who wish to communicate set up tunnels to one of these gateways.

In a similar direction, Zúquete and Frade [9] propose appropriate modifications to OpenVPN to allow for fast and seamless VPN mobility across user-provided wireless access points (APs). They assume the existence of a virtual ISP (VISP), that is, a trustworthy organization which allows users to tunnel traffic through its VPN gateways.

Although similar in spirit, our approach bears some distinct differences. While the aforementioned architectures introduce centralized entities for security management, in our case, each user is responsible for managing her own VPN gateway. Also, in [9], differentiation between hotspot owner and visitor traffic is carried out in part by the VISP and the ISP with which the micro-provider is subscribed, while in our case it is carried out autonomously at the user (micro-provider and VPN gateway owner) level. Finally, we focus on user-centric VoIP, with the aim of exposing the capabilities and performance limitations of a VoIP scheme built purely on user Wi-Fi equipment.

Security vs. Performance

VPN tunneling is known to be computationally expensive and to incur traffic overhead. Under heavy load, these factors can cause significant delays and introduce packet loss, which negatively affect user experience. These problems can become more severe when using low-end user equipment.

We thus carried out testbed measurements to quantify the effect of potential quality degradation components:

A user-centric scheme for collecting radio environment information at client spots is necessary and would offer significant advantages compared to infrastructure-centric schemes, where monitoring is carried out solely by APs at their fixed locations.

- Congestion at the Wi-Fi MAC layer
- VPN space overhead due to packet expansion caused by additional security-related protocol headers
- CPU overhead imposed due to cryptographic operations

We emulated our user-centric VoIP architecture in a testbed with two off-the-shelf Linksys WRT54GL Wi-Fi routers (W1, W2) running the OpenWRT Linux distribution and two laptop computers acting as user terminals (U1, U2), each connected to one of the two routers using IEEE 802.11g at 54 Mb/s. We emulated parallel VoIP calls between U1 and U2 by initiating bidirectional RTP/UDP streams with the traffic characteristics of the popular G.729a audio codec (60-byte IP packets, each carrying 20 bytes of audio payload; 50 packets/s), and measured end-to-end delays, jitter and packet loss ratios, which we translated to VoIP QoE estimates as specified by the International Telecommunication Union Telecommunication Standards Sector (ITU-T) E-model [10]. VoIP streams followed the path $U1 \leftrightarrow W1 \leftrightarrow W2 \leftrightarrow U_2$, and $W1$ was connected to $W2$ over 100 Mb/s Ethernet.

All functionality (wireless access, VPN management) was built into home user equipment, which is what we expect in a practical deployment of our scheme. We operated VPN gateways at $W1$ and $W2$ and set up VPN tunnels between $U1$ and $W1$, and $U2$ and $W2$. We used OpenVPN, a popular solution for implementing tunnels, which, compared to alternatives such as L2TP/IPsec, is simpler to configure, less complex protocol-wise, and has smaller per packet space overhead. Note that the connection between the two routers is unencrypted. Thus, each packet is encrypted and decrypted twice. We used AES-CBC (128-bit) for encryption, certificate-based SSL/TLS session authentication and key exchange, and HMAC-SHA1 packet authentication.

Our experimental study has shown that the performance of such a scheme is *CPU-bound*: In this setup, at most eight concurrent secure VoIP sessions of acceptable QoE are supported, and this performance limit is mainly due to the CPU-expensive encryption/decryption operations on the low-end Wi-Fi router processor.[2]

The number of concurrent acceptable-quality calls, considering that Wi-Fi router CPUs are the performance bottleneck, is a promising result. Further optimizations, such as silence suppression via voice activity detection which would reduce cryptographic load, dedicated cryptographic hardware inside Wi-Fi devices, and/or expected CPU power improvements would improve VoIP capacity.

CPU-related optimizations aside, there is another security-performance trade-off to be addressed by future studies, related to traffic redirection through home VPN gateways. On one hand, in a tunnel-based design, the user's trust anchor is his home network, and it is natural to redirect his traffic through it. On the other hand, the end-to-end path can thus become suboptimal. Tackling this issue is not straightforward. Performance studies are needed as a first step in order to quantify the potential overhead for both communicating users, but also for the network as a whole. Second, potential improvements would necessitate the involvement of the network provider, which would imply an additional level of user-ISP cooperation. Thus, new assumptions about the networking environment should be made, and the incentives of both users and operators should be taken into account.

User Feedback for Optimized Wireless Network Management

Motivation and Challenges

Real-time communication services on top of UCNs are particularly sensitive to poor signal conditions due to interference at the wireless medium that can lead to further delays and packet losses. Increased WLAN density and unlicensed spectrum scarcity, though, have aggravated the interference issue and call for sophisticated spectrum management.

Acquiring feedback as to how users perceive wireless coverage and interference becomes critical. Thus, a user-centric scheme for collecting radio environment information at client spots is necessary and would offer significant advantages over infrastructure-centric schemes, where monitoring is carried out solely by APs at their fixed locations. This way, users can have an active role in optimizing the operation of the wireless networks they access.

Focusing on wireless deployments such as UCNs mediated by community operators (e.g., Fon) or campus WLANs, where user registration and accounting are centrally controlled, we propose an architecture for collecting wireless topology/coverage information, where the task of monitoring is crowdsourced to roaming users.

However, the anticipated topology discovery accuracy improvements and cost savings due to exploiting user reports, instead of deploying additional measurement equipment and/or performing time and resource-consuming measurement campaigns, come with some security issues: Users should not be considered trustworthy, and their feedback should be carefully evaluated since they could engage in fraudulent reporting.

For this type of wireless deployment, we counter such reporting attacks by consensus-based schemes, effectively filtering fraudulent information.

Architecture

In our architecture (Fig. 3), we utilize standards-based technologies for user authentication and reporting. In particular, we apply a Radius-based scheme for authentication, authorization, and accounting, as specified in IEEE 802.11i [11]. For collecting wireless coverage information, we use IEEE 802.11k [12], a subset of which we have implemented in the Linux wireless networking stack.[3] Our system operates in *rounds*. Periodically, a centralized report collector, which maintains a global view of the network topology, requests local topology information from registered APs. The APs, in turn, send a special *beacon request* IEEE 802.11k message to authenticated clients, who respond with *beacon reports*. A beacon report is a list with information about each Wi-Fi cell in

[2] We verified, both via experiments and by a simple analytic model [7], that without VPN, 30 calls would be supported, while by disabling encryption (but keeping the OpenVPN packet structure to quantify the space overhead), this number reaches 21 calls.

[3] http://mm.aueb.gr/~pfrag/software.

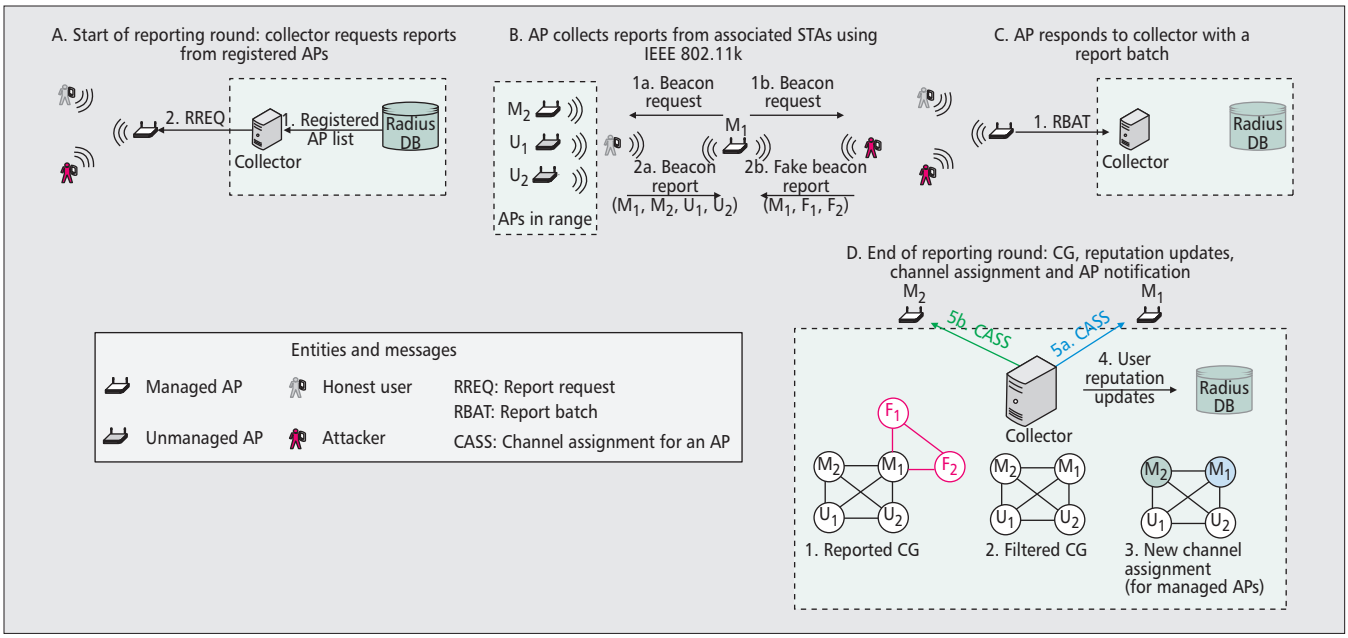


Figure 3. An architecture for crowdsourced Wi-Fi topology discovery.

the user's range. Each AP also submits its own (trustworthy) report to the collector, along with the beacon reports by users. The collector then builds a new view of the wireless topology after evaluating reports based on user reputations[4] and performing appropriate filtering to remove information it considers invalid. It can then use its up-to-date topology information as input to, say, a channel assignment algorithm, and notify registered APs about the channel they should switch to in order to minimize interference.

Model, Attacks, and Countermeasures

We model wireless topology as a weighted undirected coverage graph (CG), where vertices represent APs, and an edge between two APs denotes overlapping coverage between the respective Wi-Fi cells. Edges between two vertices corresponding to non-managed APs are not considered. The purpose of our scheme is to expose as many edges as possible, while filtering potentially fake edges that can result from fake reporting, faulty user equipment, or outdated coverage information.

Each time a user submits his feedback, and after filtering is performed, his *score* is calculated as the ratio of the information he submitted that eventually survived filtering. The user's reputation is then updated as an exponentially weighted moving average of this score. To improve the robustness of the system and help users bootstrap their reputations, trustworthy reports submitted by registered APs are used as a further means of evaluating user feedback. Each edge reported by an AP is considered valid by default, which adds to the scores of truthful users who also reported it. An edge's weight is equal to the sum of the reputations of all entities reporting it (an AP report weighs 1.0).

We consider the following attack: Each attacker acts *independently* and submits a beacon report with random fake AP identifiers. Without performing any filtering, this type of attack would result in fake vertices, and fake edges connecting them to existing ones. Since attackers do not coordinate to submit the same fake information, and the probability that more than one attackers report the same fake edge is assumed negligible, the weight of each fake edge is equal to the single reporter's reputation (thus, < 1.0). By filtering all edges the weight of which does not meet the unit-weight threshold, all fake information is eliminated. This comes with the cost of filtering some valid information that did not meet the threshold due to few reports. Also, a new snapshot of the topology is generated periodically; in the meantime, the topology may have changed due to user mobility. We show that in scenarios where user mobility is random, the ratio of *stale* to true information is minimal.

More details on our model, security measures, and their analytic performance evaluation can be found in [13].

Performance Benefits of a User-Centric Scheme

We present simulation results which show that even when large numbers of attackers are present, our user-centric reputation-based filtering scheme significantly outperforms an infrastructure-based one.

In our simulation settings, users move according to the random waypoint (RWP) mobility model [14] in a 2D terrain, where a number of APs are distributed following a spatial Poisson point process. We assume that the transmission and sensing ranges for APs and user devices are equal, and that a fixed ratio of APs are centrally managed, and a fixed ratio of users are always truthful, while the rest decide to attack at each reporting round uniformly at random with a fixed probability.

[4] A user's reputation is a real value in the [0, 1) range expressing her history of truthful reporting.

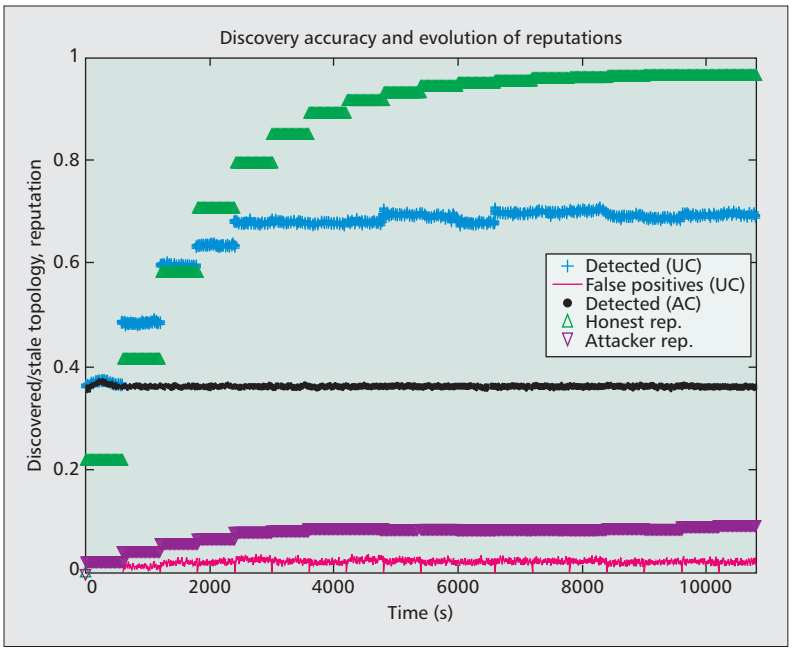


Figure 4. Wi-Fi topology discovery accuracy and evolution of reputations. The ratio of the topology (CG edges) successfully discovered in a user-centric (UC) scheme is approximately 2 that of the AP-centric (AC) one, only at the expense of 1–2.5 percent of "stale" edges caused by user mobility. Truthful users maintain very high reputations, while consistent attackers get penalized.

To select realistic values for the ratio of managed APs and AP coverage range, we performed a measurement campaign in a university building situated in a densely populated area in the center of Athens, Greece. In that setting, the density of APs was approximately 2000/km^2, less than 10 percent of which were centrally managed by the university, while the average range indoors was 30 m. We assumed a pedestrian-speed, low-mobility scenario. Reporting rounds take place every 600 s, and 50 percent of the users are potential attackers, each of whom decides to attack with 0.9 probability, independent of his choice in the previous round. The simulation starts after a warm-up period of 3600 s, necessary for the system to reach a steady state [15]. We also simulate a pure AP-centric scheme where users do not participate in the reporting process for the same simulation parameters and user mobility scenario.

Figure 4 shows a performance comparison between our reputation-based user-centric scheme and a pure AP-centric one. We present the evolution of the system's topology discovery accuracy (CG edges that get discovered), false positive rate (CG information that has become stale due to user mobility), and the evolution of the average reputation of the two categories of users (honest, potential attackers). The user-centric scheme achieves a 2× performance improvement, only at the cost of a minimal false positive rate (less than 2 percent). The reputation of honest users gradually converges to a value close to the maximum, while the reputation of consistent attackers is kept very low.

More Sophisticated Attacks

Although we limited our discussion to independent attackers, our system provides a level of resistance to coordinated attacks. The smaller the reputation of attackers, the larger the size of the group of *colluders* necessary to report the same fake information so that the latter survives filtering. Therefore, either a large group of attackers (e.g., more than 10 if the mean attacker reputation is < 0.1) should coordinate, or a user should have access to many accounts. The latter is not straightforward in UCNs with centralized user registration (e.g., Fon). Such environments also preclude attacks that involve compromised APs; in the case of a community mediator like Fon, users are assumed not to tamper with the firmware of their devices, while for deployments such as campus WLANs, APs are outside the reach of users. For more sophisticated attacks, additional measures are necessary, though. For instance, location information about managed APs (often available in managed Wi-Fi deployments) can help exclude reported edges connecting distant APs.

Conclusion

This article delves into the disruptive user-centric paradigm and its expressions in different aspects of wireless networking. We have designed and implemented architectures providing user-centric solutions for communication over open wireless access networks and for wireless network management tasks, addressing security and performance challenges and trade-offs. Respecting the principles of user-centrism, we expose performance bounds and demonstrate the benefits that can be achieved by user-centric designs, and the need for appropriate protection mechanisms.

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018," White Paper, Feb. 2014, http://goo.gl/ULXROo.
[2] J. Howe, "The Rise of Crowdsourcing," *Wired*, vol. 14, no. 6, June 2006, http://www.wired.com/wired/archive/14.06/crowds.html.
[3] P. A. Frangoudis, G. C. Polyzos, and V. P. Kemerlis, "Wireless Community Networks: An Alternative Approach for Broadband Nomadic Network Access," *IEEE Commun. Mag.*, vol. 49, no. 5, May 2011, pp. 206–13.
[4] E. C. Efstathiou, P. A. Frangoudis, and G. C. Polyzos, "Controlled Wi-Fi Sharing in Cities: A Decentralized Approach Relying on Indirect Reciprocity," *IEEE Trans. Mobile Comp.*, vol. 9, no. 8, Aug. 2010, pp. 1147–60.
[5] A. Förster *et al.*, "On Context Awareness and Social Distance in Human Mobility Traces," *Proc. 3rd ACM Int'l. Wksp. Mobile Opportunistic Networks*, 2012.
[6] B. Han *et al.*, "Mobile Data Offloading Through Opportunistic Communications and Social Participation," *IEEE Trans. Mobile Comp.*, vol. 11, no. 5, May 2012, pp. 821–34.
[7] P. A. Frangoudis and G. C. Polyzos, "On the Performance of Secure User-Centric VoIP Communication," *Computer Networks*, vol. 70, Sept. 2014, pp. 330–44.
[8] N. Sastry, J. Crowcroft, and K. Sollins, "Architecting Citywide Ubiquitous Wi-Fi Access," *Proc. ACM HotNets VI*, 2007.
[9] A. Zúquete and C. Frade, "Fast VPN mobility across Wi-Fi hotspots," *Proc. 2nd Int'l. Wksp. Security and Communication Networks*, May 2010, pp. 1–7.
[10] ITU-T Rec. G.107, "The E-model, A Computational Model for Use in Transmission Planning," ITU-T SG 12, http://www.itu.int/rec/T-REC-G.107
[11] IEEE 802.11i-2004, "IEEE Standard for Information Technology — Telecommunications and Information

Exchange Between Systems — Local and Metropolitan Area Networks — Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Medium Access Control (MAC) and Security Enhancements," IEEE 802.11 WG, 2004.
[12] IEEE 802.11k-2008, "IEEE Standard for Information Technology — Telecommunications and Information Exchange between Systems — Local and Metropolitan Area Networks — Specific Rrequirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 1: Radio Resource Measurement of Wireless LANs," IEEE 802-11 WG, June 2008.
[13] P. A. Frangoudis, *User-Centrism in Wireless Networking*, Ph.D. dissertation, Athens Univ. Economics and Business, 2012.
[14] C. Bettstetter, H. Hartenstein, and X. Pérez-Costa, "Stochastic Properties of the Random Waypoint Mobility Model," *Wireless Networks*, vol. 10, no. 5, Sept. 2004, pp. 555–67.
[15] J. Yoon, M. Liu, and B. Noble, "Random Waypoint Considered Harmful," *Proc. IEEE INFOCOM 2003*, Mar. 2003, pp. 1312–21.

BIOGRAPHIES

PANTELIS A. FRANGOUDIS (pantelis.frangoudis@inria.fr) received his Ph.D. (2012) in computer Sscience from the Department of Informatics, Athens University of Economics and Business (AUEB). He holds a B.Sc. (2003) and an M.Sc. (2005) in computer science from the same department. Currently, he is a post-doctoral researcher at INRIA Rennes-Bretagne Atlantique. His research interests include wireless networking, Internet multimedia, and network security.

GEORGE C. POLYZOS (polyzos@aueb.gr) is a computer science professor at AUEB, and founded and leads the Mobile Multimedia Laboratory. He was a CSE pProfessor at the University of California at San Diego, Computer Systems Laboratory co-director, Center for Wireless Communications Steering Committee member, and San Diego Supercomputer Center Senior Fellow. He obtained his electrical engineering Diploma from the National Technical University of Athens, and his M.Sc. in electrical engineering and Ph.D. in computer science from the University of Toronto. His current research interests include Internet architecture and protocols, mobile multimedia communications, wireless networks, ubiquitous computing, and network security.

DISASTER RESILIENCE IN COMMUNICATION NETWORKS: PART 2




Michele Nogueira *Piotr Chołda* *Deep Medhi* *Robert Doverspike*

We experience society's growing dependence on electronic communication networks in every aspect of our lives. With this comes the expectation that communication networks are readily available all the time. Networking protocols are designed to address some simple failures, such as when a packet is dropped, a retransmission occurs, or the size of the transmission window is adjusted to accommodate congestion. Similarly, routing protocols have the functionality to route around a failure. That is, communications networks have certain built-in resilience for certain specific types of failure situations. Furthermore, networks can be designed with backup paths and capacity to protect against a failure as part of critical infrastructure protection.

A disaster in a communication network is generally understood as a massive set of failures that can affect performance to the point where the degradation appreciably affects our lives. It should be noted that not all disasters are visible to end users. A disaster such as a hurricane, which encompasses a geographic area, can severely affect the communication network in the geographic area and beyond; in such a case, end users may be aware of the disaster through news reports. On the other hand, if a major disaster occurs at a large-scale data center, users from a very wide geographical region may notice degraded performance, but may not be aware of the actual event that occurred. Examples are over-the-top video provided by data centers that experience failures or with the peering locations that interface these data centers with the Internet. Millions of video customers have been affected over large areas of a country in such cases, but no environmental disaster is present. As another example, a software attack that cripples the network is a virtual disaster that will not be directly visible to end users.

It is infeasible and impractical to design and deploy hardened structures, equipment, and transmission facilities that never fail and can withstand any disaster. Therefore, the approach is to design and provide mechanisms that can recover and react to disasters. Thus, the scope of disaster resilience is that a network recovers from a disaster with an acceptable level of performance by a set of mechanisms. Such mechanisms can be either proactive, reactive, or a combination thereof. Usually, proactive approaches include redundancy in a cost-effective manner; hence, the network is sufficiently reliable to address a failure or an attack. In the case of reactive approaches, the network may react by rerouting through backup capacity, or, in some cases, by rapidly deploying ad hoc networking capability. Thus, reactive approaches may include emergency communication mechanisms during and after a disaster. The latter are an emerging and exciting approach that we have highlighted in this Feature Topic.

While much research in the past few decades focused on network resiliency, most of this work limited itself to isolated or very localized failures. When a disaster occurs, the ability of the network to recover to a reasonably acceptable performance level is a challenging problem. Thus, this Feature Topic sought submissions that covered the topic of disaster resilience from a broader perspective than had ever been attempted in the past.

The Feature Topic on Disaster Resilience in Communication Networks attracted 72 submissions. The review process resulted in identifying nine papers for publication. The first part of the Feature Topic was published in October 2014 and included four articles. The second part, in this issue, contains five articles. These five articles not only address the problems of disaster resilience in general, but also propose original and new mechanisms to support communications in case of a massive disaster.

We briefly summarize below the articles included in this issue.

The first article, "Network Adaptability to Disaster Disruptions by Exploiting Degraded-Service Tolerance" by S. Sedef Savas, M. Farhan Habib, Massimo Tornatore, Ferhat Dikbiyik, and Biswanath Mukherjee, considers a disas-

ter recovery system that assumes service differentiation (i.e., various levels of the quality provided to the clients). The authors elaborate on methods of provisioning diverse service classes from this viewpoint. According to their approach, a high level of network adaptability is feasible when it is extremely needed, which is in catastrophic events.

The second article, "Enabling Disaster Resilient 4G Mobile Communication Networks" by Karina Gomez, Leonardo Goratti, Tinku Rasheed, and Laurent Reynaud, elaborates on the virtualization of a mobile network. The virtualization is treated as a tool supporting flexibility in the reaction to disaster events and avoiding problems with the infrastructure provider. Additionally, the authors propose a distributed device-to-device communication scheme inspired by resilient tactical networking enabling the provision of high-quality transmission in the vicinity of the areas affected by catastrophes, fires, floods, and so on.

The third article, "EmergeNet: Robust, Rapidly Deployable Cellular Networks" by Daniel Iland and Elizabeth M. Belding, describes a concept of a small-scale portable cellular network that can be deployed very quickly in case of necessity. The network is designed to enable the use of basic services (e.g., messaging and voice calls based on Skype), bypassing commercial networks. Due to this idea, the network can be used by anybody in the disaster area, no matter whether or not the user is subscribed. Only a GSM phone is necessary.

The fourth article, "Exploiting the Use of unmanned Aerial Vehicles to provide Resilience in Wireless Sensor Networks" by Jó Ueyama, Heitor Freitas, Bruno S. Faiçal, Geraldo P. R. Filho, Pedro Fini, Gustavo Pessin, Pedro H. Gomes, and Leandro A. Villas, is the only article strictly focusing on communications that does not support human interaction directly, as it deals with the networks of sensors. The covered system involves relays carried by unmanned aerial vehicles, presenting a very original and promising contribution supported by broad experimental studies.

The fifth article, "Network Virtualization for Disaster Resilience of Cloud Services" by Isil Burcu Barla Harter, Dominic A. Schupke, Marco Hoffmann, and Georg Carle, presents a system establishing resilient virtual networks to be configured automatically when recovery of data flows carried inside clouds is awaited. Specifically, the authors consider the problem of choosing a proper technological layer in which recovery should be provided depending on the type of failure to bypass.

Assembling a Feature Topic on this challenging and important topic was truly a rewarding experience. We would like to thank again all the authors for their contributions to this Feature Topic and all the reviewers for volunteering their valuable time, helping to motivate the authors to make their contributions better and better.

BIOGRAPHIES

MICHELE NOGUEIRA is a professor of computer science at the Federal University of Paraná, where she has been since 2010. She received her doctorate in computer science from the Université Pierre et Marie Curie — Sorbonne Universités, Laboratoire d'Informatique de Paris VI (LIP6), in 2009. She was a visiting professor at Université Paul Sabatier and a visiting researcher at Georgia Institute of Technology in 2013 and 2009, respectively. Her research interests include wireless networks, security, and dependability. She has worked on providing resilience to self-organized and wireless networks by adaptive and opportunistic technologies such as cognitive radio. She was one of the pioneers in addressing survivability issues in self-organized wireless networks, the article "A Survey of Survivability in Mobile Ad Hoc Networks" being one of her prominent scientific contributions. She was a recipient of academic scholarships from the Brazilian Government during her undergraduate and graduate years, and of international grants such as from the ACM SIGCOMM Geodiversity program. She is also an Associate Technical Editor for *IEEE Communications Magazine* and the *Journal of Network and Systems Management*.

PIOTR CHOŁDA obtained a doctorate in telecommunications in 2006 from AGH University of Science and Technology. He then joined the Department of Telecommunications there, and now works as an assistant professor. He specializes in design of computer and communications networks. Recently, he has focused on risk-based communications networking. He is the coauthor of 16 refereed journal papers and three conference tutorials. He was a Technical Program Committee (TPC) Co-Chair of Communications QoS, Reliability and Modeling Symposium at ICC 2011, and NOMS 2014. He is a member of the Editorial Board for *IEEE Communications Surveys & Tutorials* and Editor the Book Reviews column in *IEEE Communications Magazine*.

DEEP MEDHI is a curators' professor in the Department of Computer Science and Electrical Engineering at the University of Missouri-Kansas City (UMKC). He received his B.Sc. in mathematics from Cotton College, Gauhati University, India; his M.Sc. in mathematics from the University of Delhi, India; and his Ph.D. in computer sciences from the University of Wisconsin-Madison. Prior to joining UMKC in 1989, he was a member of technical staff at AT&T Bell Laboratories. He was an invited visiting professor at the Technical University of Denmark, a visiting research fellow at Lund Institute of Technology, Sweden, a research visitor at the University of Campinas, Brazil, under the Brazilian Science Mobility Program, and served as a Fulbright Senior Specialist. He is currently Editor-in-Chief of Springer's *Journal of Network and Systems Management*, and is on the Editorial Boards of *IEEE/ACM Transactions on Networking*, *IEEE Transactions on Network and Service Management*, and *IEEE Communications Surveys & Tutorials*. He was an Associate Technical Editor of *IEEE Communications Magazine* from 2006 to 2010. His recent IEEE activities include serving as a TPC Co-Chair of IEEE NOMS 2010, General Chair of IEEE CloudNet 2013, and Vice-Chair of the Committee on Network Operations and Management (CNOM). He is coauthor of the books *Routing, Flow, and Capacity Design in Communication and Computer Networks* (Morgan Kaufmann/Elsevier, 2004) and *Network Routing: Algorithms, Protocols, and Architectures* (Morgan Kaufmann/Elsevier, 2007).

ROBERT DOVERSPIKE [F] received his undergraduate degree from the University of Colorado, and his Master's and Ph.D. degrees from Rensselaer Polytechnic Institute (RPI). He began his career with Bell Labs and, upon divestiture of the Bell System, went to Bellcore (later called Telcordia). He returned to AT&T Labs (Research), where he is now executive director of network evolution research. He has made extensive contributions to the field of optimization of multi-layered transmission and switching networks, and pioneered the concept of packet transport in metro and long distance networks. He also pioneered work in spearheading the deployment of new architectures for transport and IP networks, network restoration, and integrated network management of IP-over-optical-layer networks and software defined networking. He has over 1500 citations of his books and articles over diverse areas/publications such as telecommunications, optical networking, mathematical programming, *IEEE Communications Magazine*, IEEE Communications Society, operations research, applied probability, and network management. He holds many professional leadership positions and awards, such as an INFORMS Fellow, member of the Optical Society of America, co-founder of the INFORMS Technical Section on Telecommunications, OFC Steering Committee, DRCN Steering Committee, and Associate Editor for the *Journal of Optical Communications and Networking*.

Network Adaptability to Disaster Disruptions by Exploiting Degraded-Service Tolerance

S. Sedef Savas, M. Farhan Habib, Massimo Tornatore, Ferhat Dikbiyik, and Biswanath Mukherjee

ABSTRACT

The rapid increase in network traffic with new bandwidth-hungry applications such as cloud computing and telemedicine makes disaster survivability a crucial concern as the data (and revenue) loss caused by large-scale correlated cascading failures can be very high. To alleviate their impact, new measures should be taken since the nature of the network changes dramatically as available resources decrease during disasters. We develop a metric, called degraded-service tolerance, which can reduce protection cost and network disruption, and support maximal carried traffic in case of disasters. Degraded-service-tolerant connections can be admitted and recovered with reduced bandwidth under resource crunch. Our scheme re-assigns resources among connections by leveraging their degraded-service tolerance. A case study shows how our proposal can be applied to boost network performance during the resource crunch following a disaster.

INTRODUCTION

Telecom networks are exposed to many threats such as malicious attacks, equipment failures, human errors (e.g., misconfigurations), and large-scale disasters, both human-made (e.g., due to weapons of mass destruction, WMD, attacks) and natural. Disasters represent a challenging threat for our networks as they affect large geographical areas, and can trigger correlated and/or cascading failures of multiple network elements, resulting in huge data loss and disruptions in network connectivity. The 2011 Japan Earthquake/Tsunami and 2012 Hurricane Sandy [1] are two recent disasters that have deprived people of essential network services and severely hindered rescue operations for weeks. Thus, disaster-aware provisioning schemes (e.g., routing around risky disaster areas) have been proposed. The emergence of bandwidth-hungry applications has led to rapid growth in the volume of the data traffic in our networks, making cost-effective survivability methods against disasters even more crucial.

Today's networks support diverse services, from cloud computing and video streaming to traditional ones (HTTP, VoIP, etc.). Services have different requirements (e.g., delay/latency tolerance and bandwidth) and characteristics (e.g., importance and revenue generation). With such heterogeneity, using the same protection policies for all services can result in suboptimal solutions. Thus, we consider the different tolerances of services (i.e., degraded-service tolerance) to develop fault-tolerant (survivable) methods that can sustain an acceptable level of service even when disasters occur. Some services are sensitive to the amount of capacity provided, while others (e.g., video streaming or file transfers) can operate with reduced bandwidth. *Degraded service* refers to a reduced amount of resource allocation for a service vs. its normal operational requirement. Thus, even with degraded service, some services can still achieve lower but acceptable quality.

The degraded-service concept has been investigated for survivable service provisioning schemes against large-scale disasters [2]. Providing 100 percent protection against disasters (by routing them via primary and backup paths) would require massive and economically unsustainable bandwidth overprovisioning, as disasters are difficult to predict and statistically rare, and may create large-scale failures. Some researchers have shown that providing protection for a portion of the requested bandwidth, or *partial protection*, can alleviate the extensive resource usage of full protection schemes [2]. The fraction of the requested bandwidth that will be guaranteed, even under failures, is determined by the degraded-service tolerance of services, generally stated in a service level agreement (SLA).

A limitation of prior studies exploiting degraded service is that they do not adapt their resource allocation based on the network state, which leads to suboptimal network utilization in the presence of disasters. They exploit degraded-

S. Sedef Savas, M. Farhan Habib, and Biswanath Mukherjee are with the University of California, Davis.

Massimo Tornatore is with the University of California, Davis, and Politecnico di Milano.

Ferhat Dikbiyik is with Sakarya University.

This work has been supported by the Defense Threat Reduction Agency (DTRA) Grant HDTRA1-10-1-0011.

A preliminary version of this work was presented at OFC 2014. This extended work includes detailed analysis specific to disaster scenarios.

0163-6804/14/$25.00 © 2014 IEEE

service tolerance of connections only during admission to provide them partial protection so that when a disaster occurs, the connection will continue to operate at its minimum required service level [1, 3]; hence, when a disaster occurs, already accepted (and unaffected) connections are not considered by these studies for network resource optimization.

Networks may experience *resource crunch*, that is, an undesired reduction of network capacity due to disaster failures. Since disasters are rare, resource crunch is an unusual situation, so different measures should be taken to get through disasters with minimal damage. There are two major contributions of this article besides exploiting degraded-service tolerance as partial protection schemes do. First, our disaster-aware provisioning considers the decision process in the aftermath of disasters separate from the provisioning step, as it also considers the degraded-service tolerance of incoming connections. Second, during disasters, we allow provisioned connections to also be part of the resource allocation optimization. To offer the affected connections acceptable levels of service, unaffected connections can be degraded, rerouted, or even halted depending on their importance or profitability characteristics.

In this article, we first present an overview of existing disaster-aware service provisioning schemes that exploit degraded-service tolerance. Then we describe our disaster-aware adaptive resource allocation scheme. Finally, we illustrate and quantitatively compare these mechanisms using a case study in wavelength-division multiplexing (WDM) mesh networks. Note that our techniques are applicable to general mesh networks as well.

Overview of Disaster-Aware Provisioning Schemes

Disaster repercussions (e.g., disconnections, data loss, and service disruptions) can be minimized using protection schemes (preconfigured before disasters) and restoration schemes (reactive, after disasters).

Disaster-Aware Provisioning Schemes

Some studies propose survivable provisioning to proactively alleviate the impact of disasters. They determine possible disaster zones in the network, such as *risk (hazard) maps* to highlight its vulnerable regions using interdisciplinary contributions from climatology, geology, and so on. (Figure 1a shows risky zones of the United States with a heat map against some natural disasters [4]). By utilizing risk maps, we can estimate the probability of occurrence of a disaster and probability of a network device getting damaged by this disaster. These two parameters give us the risk levels of disaster zones. Figure 1b shows a seismic-hazard map of the United States with its risk levels. Locations of high-risk earthquake zones (Fig. 1c), which have different probabilities of failures, can be determined by matching a network topology with the seismic hazard map. Using these maps, network planners can develop systems that select less risky regions for routing connections; hence, the expected loss will be minimized and the network becomes better prepared to handle a disaster.

The set of links or nodes that are vulnerable to a common failure (e.g., a disaster) can be represented as a shared risk group (SRG) [1]. The most prevalent protection strategy against disasters is to route connections over disaster-zone-disjoint (i.e., SRG-disjoint) primary and backup paths (or using multiple primary paths, e.g., multi-path provisioning). However, fully protecting primary paths with backups requires extensive resource usage, especially for multiple failures (as in disasters).

Some services may accept a reduced level of bandwidth during failures, depending on their characteristics. For services that can tolerate reduced bandwidth, network operators may offer partial protection, possibly at lower cost. The partial protection guarantee is determined by the connection's degraded-service tolerance.

Multipath routing (i.e., multiplexing a connection over multiple paths) is another scheme for providing partial protection [6]. For a multipath-routed service, even if some paths are down or overloaded, other paths may provide the required degraded service. Thus, some SLAs for partial protection can be satisfied without any redundant resource allocation.

The shifting paradigm toward cloud computing is creating new opportunities for optimizing disaster-aware network design. Contents in cloud systems can be replicated at multiple servers/data centers from which users can be served. New service models are introduced such as *anycasting* (providing service from any of the data centers that host the requested service) and *manycasting* (providing service from a subset of the data centers). These models can be exploited, for example, for file transfer and media streaming to enhance the resilience of cloud services. Resilience against destination node failures is very crucial due to cloud services and data centers hosting content. These schemes (shown in Fig. 2) are resilient against destination-node failures (since the paths connect to disaster-zone-disjoint destination nodes, so the service will not be disrupted if such a node fails due to a disaster).

Figure 2 shows examples of a degraded-service-aware single-primary path using anycasting and multipath provisioning using manycasting [2], both of which guarantee a minimum tolerable bandwidth of 60 percent of the required bandwidth under normal operation even under disaster scenarios. Backup-path protection using anycasting is shown in Fig. 2a, where a 1 Gb/s connection is partially protected by a backup path with 0.6 Gb/s capacity. Reference [7] used inverse multiplexing over multiple paths (the least risky ones) to provision bandwidth for services distributed over multiple servers/data centers with manycasting (Fig. 2b). Also, it ensures degraded service (vs. no service at all) after a failure without using extra resources since it uses multipath routing. For instance, in Fig. 2b, during normal operation, the customer receives 1 Gb/s service, which is multiplexed over three paths (one with 0.4 Gb/s and two with 0.3 Gb/s) destined to different data centers; and any pre-

For a multi-path-routed service, even if some paths are down or overloaded, other paths may provide the required degraded service. Thus, some SLAs for partial protection can be satisfied without any redundant resource allocation.

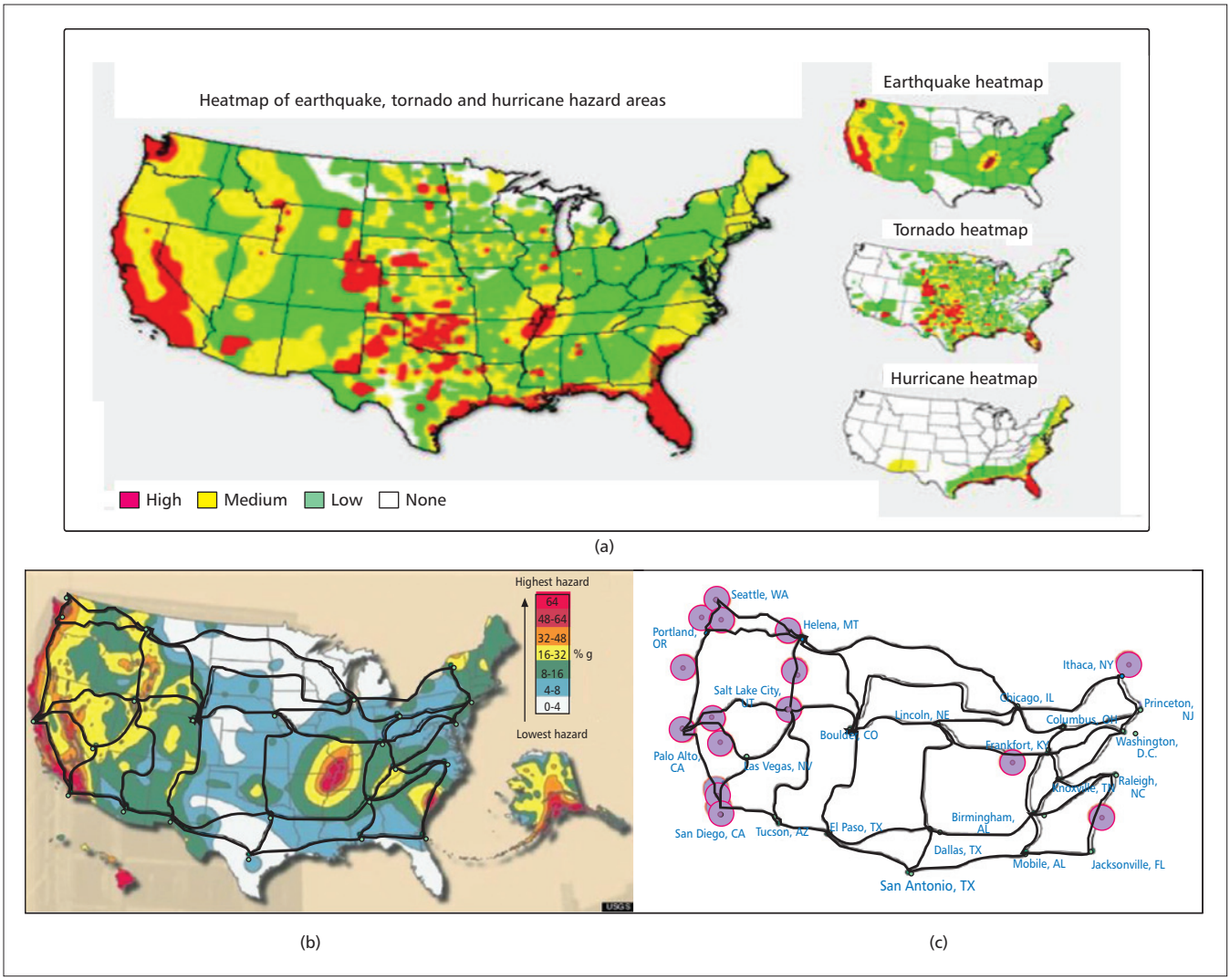


Figure 1. Exploitation of hazard maps to determine disaster zones: a) natural disaster risk map (Credit: U.S. Geological Survey); b) earthquake risk map; c) earthquake zones (shown in circles).

dicted disaster in the figure affects only one path, so the guaranteed bandwidth is at least 0.6 Gb/s.

Restoration/Reprovisioning Schemes

Despite the above measures, it is not always possible to avoid all disaster zones and provide protection to all services for all disaster scenarios; moreover, unforeseen attacks and disasters may occur. Therefore, taking actions in the aftermath of a disaster should also be considered. Reprovisioning is a reactive approach where network resources are re-allocated for existing connections. If an unpredicted disaster occurs, restoration schemes can be used to preserve the targeted quality of service (QoS) level or to ensure graceful degradation using the remaining resources in the undamaged parts of the network. Since usually only some parts of the network are damaged, the unused capacity available in the network's remaining parts can be used to reprovision the disrupted connections. Note that during restoration, secondary failures such as aftershocks following an earthquake should be considered as they are predictable with good accuracy after the primary failure. Reprovisioning schemes are robust against different types of failures as they adapt the network according to its current state but they do not give restoration guarantee for the disrupted services as provisioning schemes do. Full-service restoration schemes are as costly as full protection schemes, and reduce restoration chances due to high bandwidth requirement in a limited-resource environment. Partial-bandwidth restoration [1] (exploiting degraded-service tolerance) may be a good option as it requires fewer resources and hence increases restoration probability. Although resource consumption is low, restoration schemes are not favored for live traffic (e.g., rerouting primary paths), as they cause service disruptions due to reprovisioning of network resources.

To handle the unusual resource crunch problem caused by disaster failures, besides exploiting degraded-service tolerance just for partial-protection purposes as the above works do, we propose to exploit it further to perform degraded-service-aware resource re/allocation in case of disasters. Our solution is applicable for

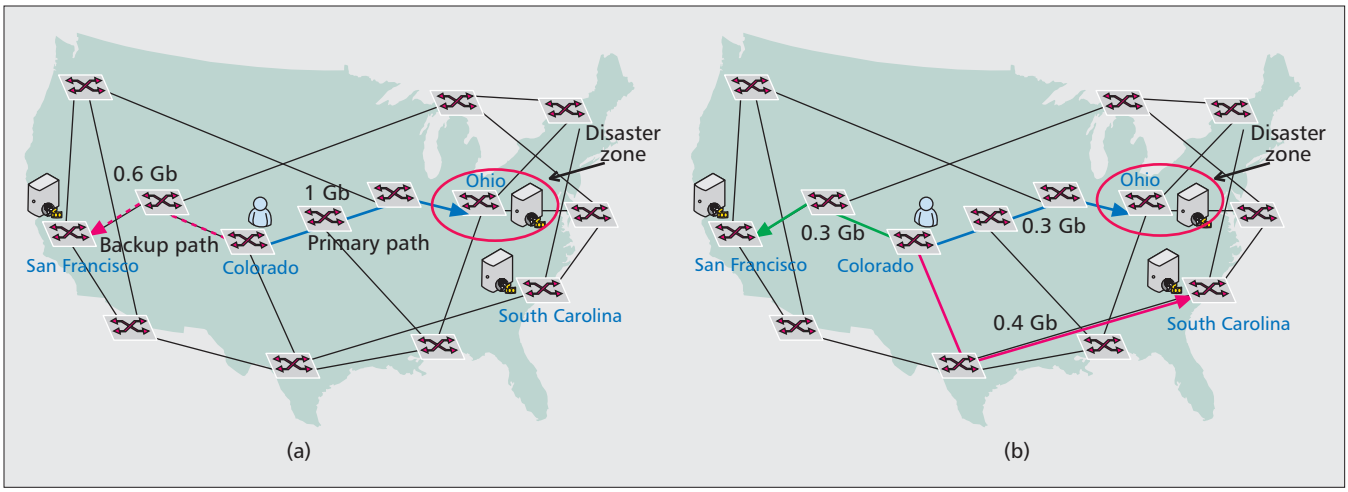


Figure 2. Disaster-aware provisioning schemes in cloud networks: a) anycasting with partial protection; b) manycasting (multi-path) with partial protection.

all networks with service differentiation functionality, such as optical WDM networks, IntServ, DiffServ over IP networks, multiprotocol label switching (MPLS) on layer 2 networks, and software defined networking (SDN)-enabled networks [8].

The Proposed Approach: Degradation-Aware Adaptive Resource Allocation

The above solutions perform static resource allocation; for example, once admitted, no further actions are taken on a connection unless there is a fault that affects it. Such non-adaptive approaches result in suboptimal solutions.

We propose and study the characteristics of a novel resource allocation framework (service provisioning and reprovisioning) to enable the network to be robust to disasters by adaptively responding to changes in the network state. This is a new approach to (re)distribute resources among existing connections. Our work can coexist with prior disaster-aware provisioning schemes. We aim to minimize blocking rate (unadmitted connection ratio over all requests), dropping rate (losing a connection), and disruption rate (rate of reprovisioning of connections) during disasters. Also, we aim to provide the best service possible to connections with remaining network resources by rearranging the resource allocation if needed.

We exploit the degraded-service concept to combat disasters as follows:

1. Accept degraded services during call admission to increase service acceptance rate, and if necessary, degrade existing connections
2. Degrade existing connections to reduce dropping rate
3. Apply an upgrade process to restore degraded connections to full service whenever possible

Figure 3 shows our proposal in three steps: *provisioning*, *recovery*, and *upgrade*. The *provisioning step* is applied only when a disaster occurs; otherwise, disaster-aware provisioning with full service is used. Specifically, we deal with the limitation of available resources caused by a disaster during the admission process by accepting connections with degraded service and/or adapting the network by degrading existing connections to release some resources for incoming connections. *The recovery step* kicks in when some connections get disrupted due to a disaster failure; then we try to release some capacity to avoid them being dropped. Finally, the *upgrade step* is triggered when new resources are reactivated after the repair of some network elements, to provide the best service to connections within available network resources.

Our approach may increase management complexity under faulty network conditions. This increase might be considered acceptable as it represents the cost to reduce the connection dropping rate and provide the best possible service with the residual resources while not violating SLAs. Recent progress in automatic control/management solutions (e.g., SDN) can provide technological support to harness this increase in management complexity. SDN eases the network control, monitoring, and management, thus enabling more dynamic approaches such as our proposed solution. Modifying end-to-end service levels of existing connections that are not affected by the disaster requires a multi-tenant, multi-layer management scheme, which will be enabled by SDN. This multi-layer optimization together with a holistic view of the network may increase the chances of discovery of resources, which is crucial in a stressed scenario such as a disaster. This feature also improves disaster recovery time as post-disaster convergence of the network may take a large amount of time (seconds) if addressed only at a higher layer of the network such as the IP layer.

With evolving cloud and video applications, traffic patterns are becoming more dynamic. Therefore, even without disasters, the overload from sudden bursts of traffic can lead to resource crunch. To cost-effectively handle growing/bursty traffic, networks may need to be run at high uti-

The decision of what to degrade and how much to degrade will depend on the optimization objective. This decision affects the number of connections that are reprovisioned, and the degradation process can be performed by allowing either global reprovisioning or essential (local) reprovisioning.

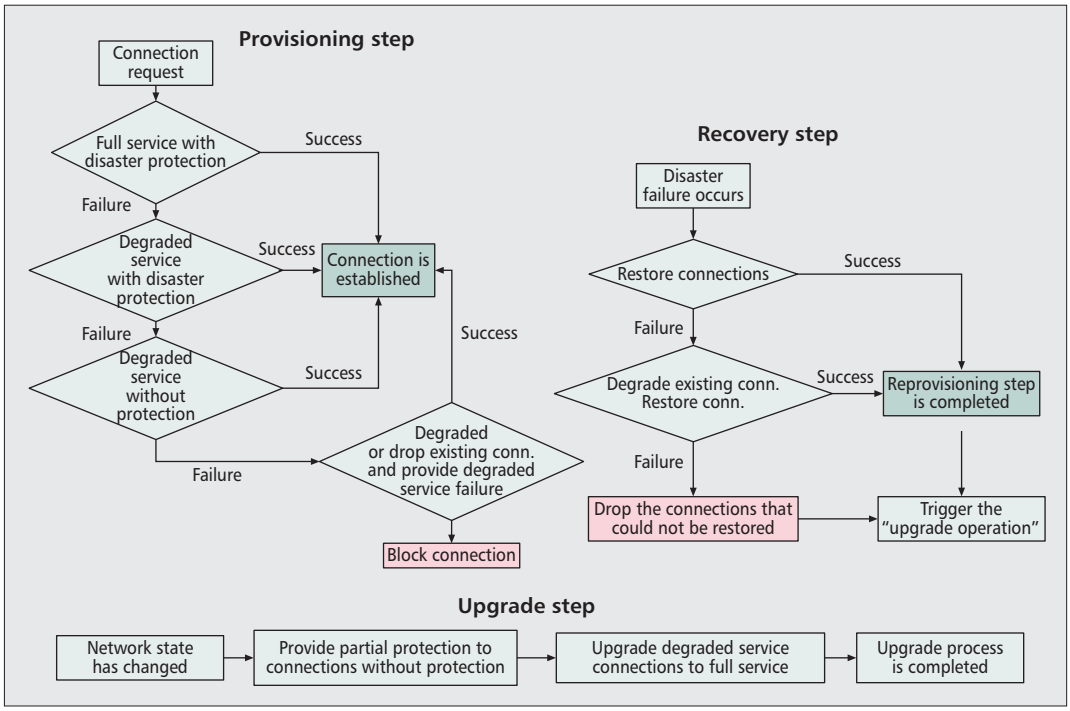


Figure 3. Operational steps.

lization [9]. These highly utilized networks are at risk of resource crunch, so our adaptive resource allocation scheme can apply to them as well.

What and How to Degrade

Degraded service is exploited to achieve effective traffic engineering with high network utilization in disasters. The decision of what to degrade and how much to degrade will depend on the optimization objective. This decision affects the number of connections that are reprovisioned, and the degradation process can be performed by allowing either global reprovisioning or essential (local) reprovisioning. For instance, as a result of a global reprovisioning approach (no restriction on the number of reprovisioned connections), better resource utilization may result in better network performance in terms of recovered connections and blocking probability. However, a high number of reprovisioning actions may cause excessive disruptions, which in turn could cause data loss due to the switching time required to reprovision the connections. There is a trade-off between better network utilization and providing uninterrupted service. Some optimization objectives to determine the level of flexibility between these two are as follows (and note that only the first objective below is used in our current study):

- *Maximize profit*: Prioritizing the low-revenue-generating services in the degradation process can maximize profit by degrading the minimum number of connections that generate low income.
- *Minimize penalties for SLA violations*: Some penalty can be applied to a service when the service level goes below its agreed QoS in SLA. There is a minimum service level that a customer can accept, and this must be satisfied.
- *Customer satisfaction*: Customers can tolerate degraded service under unusual circumstances. For better customer satisfaction, the full bandwidth requested must be provided. To minimize dissatisfaction from degraded service, the fewest connections should be degraded whenever possible.
- *Load balancing*: If some resources need to be released, while selecting connections to be degraded we need to consider their physical locations as well. As the number of degraded-service connections increases in a region, the chance of upgrading them to full service will be less than a scheme where connections to be degraded are selected by region (i.e., regions with a high density of degraded services will not be selected).

Also, the optimization objective can be a mix of some of the above objectives; this is a topic for further investigation.

Other Opportunities to Exploit Service Differentiation Characteristics

Services have different characteristics (e.g., revenue generation, importance) and different requirements (e.g., delay tolerance, availability, and bandwidth) [10]. These metrics should be exploited in degraded-service provisioning to determine which services should have reduced capacity (and by how much) under resource crunch. Below, we list some possible service differentiations:

- *Scalability*: Scalability of services is an important factor for the degradation process. Services may not operate at every bit rate between a requested bit rate and the minimum tolerable bit rate. For instance, if we assume that a streaming video is encoded at two different qualities (high and low),

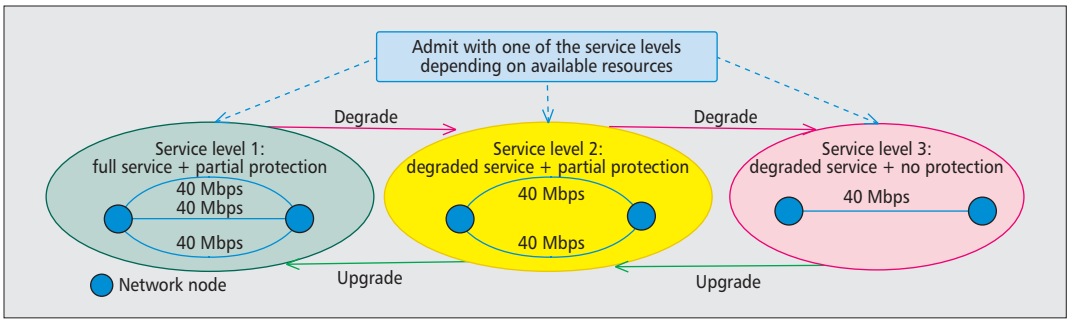


Figure 4. Possible service levels provided during admission. (The service level of a connection may be degraded during provisioning and recovery steps to free some resources for incoming or disrupted connections. The service level is only upgraded during the upgrade step.)

if the bit rate allocated to this service is between low and high bit rates, low-quality video will be served and the remaining bit rate will be wasted. Thus, this aspect should be considered.

- *Mission-critical vs. non-mission-critical*: Some services are critical such as military services. Service availability is increasingly becoming a requirement for certain mission-critical applications. Attention must be paid for these connections to reduce their disruption rate.
- *Real-time vs. data traffic*: Traffic can be classified as:
  - –Real-time (delay-sensitive, e.g., VoIP, video conferencing, web search) and
  - –Data (non-real-time traffic, which is not delay-sensitive, e.g., file transfer)

Our solution can exploit these QoS differences. While we do not reprovision delay-sensitive connections as reprovisioning causes delay, non-real-time traffic can be reprovisioned to release resources that can be utilized by other connections to recover from degradations to their normal operational state.

Degraded-Service-Aware Provisioning: A Case Study

We present a case study to show how our proposed framework can be applied to enhance resilience, resource utilization, and QoS. Our study covers two differentiated service types:

- Degraded-service-tolerant and -intolerant services
- Mission-critical and regular services

A dynamic scenario is considered where traffic/service requests arrive, hold for a while, and then depart. Also, disasters randomly hit the network, and affected resources remain unavailable until they get repaired. Disaster protection is granted to connections by multiplexing the connection over SRG-disjoint paths (multipath provisioning). In case of a multi-path-routed service, even if a path is down or overloaded, the other paths may provide the required degraded service. Thus, SLAs can be satisfied without additional resource usage or re-allocation of bandwidth among competing resources. In our study, each service level provides partial protection, except degraded-service-intolerant services (which always require full bandwidth, i.e., full protection).

We consider three service levels which can be provided to a connection during its lifetime. These three service levels can be seen in Fig. 4 (service levels 1–3 in descending order of quality) for a connection request with 120 Mb/s full service and 40 Mb/s degraded service requirement:

- *Service level 1 (full service with partial protection)*, which provides full bandwidth during normal operation and degraded service during failures
- *Service level 2 (degraded service with partial protection)*, which provides the requested degraded service during normal operation and partial protection after failures
- *Service level 3 (degraded service without protection)*, which provides the requested degraded service during normal operation and does not give any guarantee after failures

Without causing any disruption, by tearing down (degrading) or adding some paths (upgrading) of/to a connection, we can adapt the network according to its current state. Our proposed solution is not restricted by multipath provisioning and can be applied to any existing provisioning scheme.

We analyze the steps (provisioning, recovery, and upgrade) depicted in Fig. 3 by comparing three different approaches explained below. The *degraded-service* and *extreme-degraded-service schemes* are the proposed approaches, whereas the *full-service scheme* is the traditional approach. These schemes can work with any number of service levels. For illustration, we consider the above-mentioned three service levels.

Full-service scheme (FS): In FS, degraded service is not exploited during resource crunch; that is, no reallocation of network resources is allowed, and connections are only accepted with full service. This is the traditional admission approach where an incoming connection is denied if the network cannot provide full service. In the recovery step, disrupted connections are attempted for recovery to full service by reprovisioning. If not successful, they are dropped.

Degraded-service scheme (DS): At admission, connections under DS can be accepted with the best possible service level between the ranges of degraded service to full service (service levels 1–3). During recovery, we reprovision disrupted connections by gradually degrading the service level until it is satisfied. In this scheme, realloca-

Without causing any disruption, by tearing down (degrading) or adding some paths (upgrading) of/to a connection, we can adapt the network according to its current state. Our proposed solution is not restricted by multipath provisioning and can be applied to any existing provisioning scheme.

tion of network resources among existing connections to release resources is not allowed.

Extreme-degraded-service scheme (EDS): At both admission and reprovisioning, this scheme allows rearrangement in the network by relocating or degrading existing connections to release resources for incoming and disrupted connection after trying the steps in the degraded-service scheme. This scheme includes all steps shown in Fig. 3.

The upgrade step in Fig. 3 can be applied to all of the above schemes. The upgrade step is triggered after each network state change to provide the best possible service according to current network state. The network state changes whenever:

- A new request arrives.
- An existing connection terminates.
- A network failure occurs.
- A failed link/node is repaired.

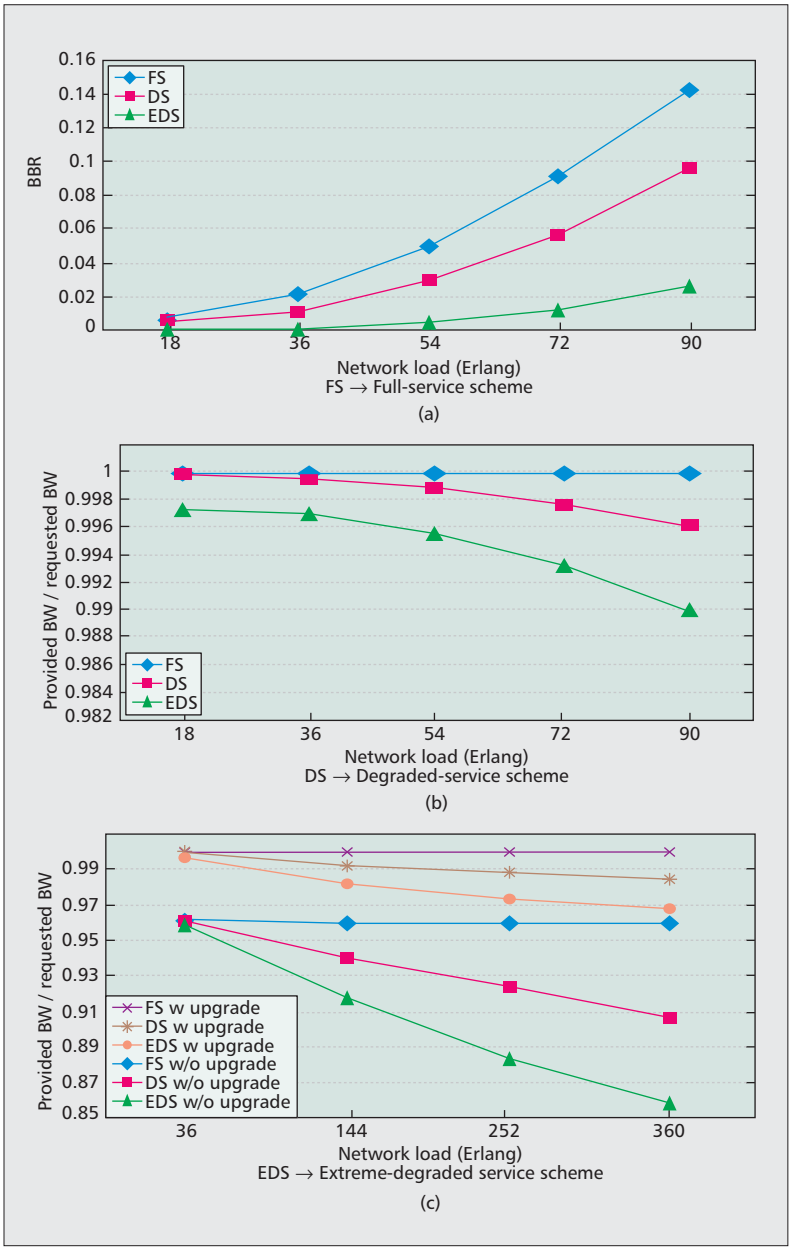


Figure 5. Comparison of the proposed schemes (DS and EDS) vs. a non-adaptive provisioning scheme (FS): a) bandwidth blocking probability (BBR); b) average bandwidth provided; c) effect of upgrade operation.

In the upgrade step, first, partial protection is provided to connections that do not have protection. Then connections that receive degraded service are upgraded to full service if possible. Even though our degraded-service schemes increase acceptance rate (at the cost of slightly decreasing average bandwidth provided), the upgrade operation alleviates the decrease in average bandwidth, which stems from the fact that some existing connections' service levels are degraded to accommodate new ones. During the upgrade step, to reduce disruptions, we only reprovision a connection while upgrading it from service level 3 or 2 to service level 1.

Case Study: Results and Evaluation

To evaluate the proposed approaches, we simulated a realistic dynamic environment. The connection arrival process is Poisson with arrival rates 0.3, 0.6, 0.9, 1.2, and 1.5 requests/min for Figs. 5a–5b; and 0.6, 2.4, 4.2, and 6 requests/min for Fig. 5c. Connection holding time follows a negative exponential distribution with a mean of 60 min. We simulate wavelength-division-multiplexed (WDM) optical networks. Each link has 10 wavelengths, and each wavelength has 10 Gb/s of capacity. The bandwidths of the connection requests follow the realistic distribution 150M: 600M: 2.5G: 10G = 40: 30: 20: 10 [2]. Degraded-service requirements of connections are uniformly distributed between 40 and 100 percent of their requested bandwidth (on average 70 percent), and connections that need the whole requested capacity at all times are intolerant to degraded service. We simulated 100,000 connection requests on a sample 24-node U.S. topology with the shown disaster scenario, with occurrence rate 2×10^{-5} disasters/min (Fig. 1c). In this study, connections arrive with the information of their connection type (either mission-critical, which is 15 percent of all connections, or regular) and degraded-service tolerance. We prioritize mission-critical connections at every decision step (e.g., we upgrade mission-critical connections first).

Figure 5a compares bandwidth-blocking ratio (BBR) (i.e., the amount of rejected bandwidth over the total requested bandwidth) of the FS scheme with our degraded-service approaches. At low loads, there is no significant difference as acceptance ratio is high, but at high loads, our degraded-service schemes outperform the traditional scheme by decreasing BBR significantly. Since our schemes are much more flexible in terms of resource allocation, we can serve more connections with the same amount of bandwidth as the traditional scheme.

Since connections may receive degraded service in DS and EDS, the average bandwidth provided by these schemes is slightly lower than in FS (Fig. 5b). Figure 5c shows the benefit of upgrade; and to observe it, the connection arrival rates for this figure are higher than the others as at low load, disrupted connections can get reprovisioned with full service. At high network loads (e.g., 360 Erlang in Fig. 5c), the proposed degraded service schemes affect average provided bandwidth excessively, as much as 15 percent decrease, when no special action is taken. Our upgrade mechanism successfully remedies this

problem by reducing the decrease in provided bandwidth from 15 percent in EDS without upgrade to 3 percent with upgrade and from 10 percent in DS without upgrade to 1 percent with upgrade (when load is 360 Erlang in Fig. 5c). Also, we observe that the acceptance and recovery rates of mission-critical connections are higher than regular connections as critical connections are prioritized.

Conclusion

Recent disasters have shown that current survivability schemes in our networks are lacking, and enhanced schemes need to be considered. By exploiting service differentiation and the degraded-service concept, we can improve the network's adaptability against disasters. We propose a method that accepts service degradation not only during failures but also during the admission process to increase service acceptance and/or availability. Also, in our proposal, some additional resources can be released by downgrading some existing connections to avoid dropping some other connections that are affected by the disaster. A case study where our proposal methods are applied to a U.S.-wide network topology shows our method's advantageous properties.

References

[1] M. F. Habib *et al.*, "Disaster Survivability in Optical Communication Networks," *Computer Commun.*, vol. 36, no. 6, Mar. 2013, pp. 630–44.

[2] S. Huang *et al.*, "A Multistate Multipath Provisioning Scheme for Differentiated Failures in Telecom Mesh Networks," *J. Lightwave Tech.*, vol. 28, no. 11, June 2010, pp. 1585–96.

[3] H. Chang, "A Multipath Routing Algorithm for Degraded-Bandwidth Services Under Availability Constraint in WDM Networks," WAINA, 2012.

[4] P. Agarwal *et al.*, "The Resilience of WDM Networks to Probabilistic Geographical Failures," *IEEE/ACM Trans. Networking*, vol. 21, no. 5, 2013, pp. 1525–38.

[5] T. L. Weems, "How Far is Far Enough," *Disaster Recovery J.*, vol. 16, no. 2, Spring 2003.

[6] W. Zhang *et al.*, "Reliable Adaptive Multipath Provisioning with Bandwidth and Differential Delay Constraints," *IEEE INFOCOM*, Mar. 2010.

[7] S. S. Savas *et al.*, "Disaster-Aware Service Provisioning by Exploiting Multipath with Manycasting in Telecom Networks," *IEEE ANTS*, Chennai, India, Dec. 2013.

[8] N. Bitar, S. Gringeri, and T. J. Xia, "Technologies and Protocols for Data Center and Cloud Networking," *IEEE Commun. Mag.*, vol. 51, no. 9, pp. 24-31, Sept. 2013.

[9] S. Jain *et al.*, "B4: Experience with A Globally-Deployed Software Defined WAN," *ACM SIGCOMM*, Aug. 2013.

[10] S. Oueslati and J. Roberts, "Method and a Device for Implicit Differentiation of Quality of Service in a Network," U.S. Patent No. 7,646,715, 12 Jan. 2010.

Biographies

S. Sedef Savas (ssavas@ucdavis.edu) is currently pursuing her Ph.D. degree in computer science from the University of California at Davis. She received her B.S. in computer science and engineering from Sabanci University, Istanbul, Turkey in 2009, and her M.S. in computer science from Koc University, Istanbul, Turkey in 2011. Her research interests include adaptability and survivability of communication networks against disasters, QoS-aware service provisioning schemes, and Software-Defined Networking (SDN).

M. Farhan Habib (mfhabib@ucdavis.edu) received the B.Sc. (Engg.) degree in computer science and engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, and the M.S. degree in computer science from the University of California, Riverside, in 2004 and 2010, respectively. He received his Ph.D. in computer science from the University of California, Davis in 2014. His research interests include network survivability, datacenter networks, network modeling, and optimization.

Massimo Tornatore (tornator@elet.polimi.it) is currently an associate professor in the Department of Electronics, Information and Bioengineering at Politecnico di Milano, where he received a Ph.D. degree in information engineering in 2006. He also holds an appointment as an adjunct associate professor in the Department of Computer Science at the University of California, Davis, where he served as a postdoc researcher in 2008 and 2009. He has co-authored more than 180 conference and journal papers and his research interests include design, protection, and energy efficiency in optical trasport and access networks and group communication security. He was a co-recipient of six Best Paper Awards from IEEE conferences.

Ferhat Dikbiyik (fdikbiyik@sakarya.edu.tr) is an assistant professor in the Department of Computer Eng. at Sakarya University, Turkey. He received his MS and Ph.D. in electrical & computer engineering at the University of California, Davis in 2009 and 2013, respectively. He received his BS degree from Istanbul University, Turkey, in 2005. His research interests include design, development, and analysis of next-generation lightwave networks, especially excess capacity management, survivability against disasters, network upgrade, and cyber security.

Biswanath Mukherjee [F] (bmukherjee@ucdavis.edu) is a distinguished professor at the University of California, Davis, where he was Chairman of Computer Science from 1997 to 2000. He received the BTech degree from the Indian Institute of Technology, Kharagpur (1980) and a Ph.D. from the University of Washington, Seattle (1987). He was General Co-Chair of the IEEE/OSA Optical Fiber Communications (OFC) Conference 2011, Technical Program Co-Chair of OFC'2009, and Technical Program Chair of the IEEE INFOCOM'96 conference. He is the editor of Springer's Optical Networks Book Series. He has served on eight journal editorial boards, most notably *IEEE/ACM Transactions on Networking* and *IEEE Network*. In addition, he has guest edited special issues of the *Proceedings of the IEEE*, *IEEE/OSA Journal of Lightwave Technology*, *IEEE Journal on Selected Areas in Communications*, and *IEEE Communications*. He has supervised 61 Ph.D.s to completion and currently mentors 18 advisees, mainly Ph.D. students. He is the winner of the 2004 Distinguished Graduate Mentoring Award and the 2009 College of Engineering Outstanding Senior Faculty Award at UC Davis. He is co-winner of Optical Networking Symposium Best Paper Awards at IEEE Globecom 2007 and 2008. He is author of the graduate-level textbook *Optical WDM Networks* (Springer, January 2006). He served a five-year term on the Board of Directors of IPLocks, a Silicon Valley startup company. He has served on the Technical Advisory Board of several startup companies, including Teknovus (acquired by Broadcom).

By exploiting service differentiation and the degraded-service concept, we can improve the network's adaptability against disasters. We proposed a method that accepts service degradation not only during failures but also during admission process to increase service acceptance and/or availability.

Enabling Disaster-Resilient 4G Mobile Communication Networks

Karina Gomez, Leonardo Goratti, Tinku Rasheed, and Laurent Reynaud

ABSTRACT

4G Long Term Evolution is the cellular technology expected to outperform previous generations and to some extent revolutionize the experience of users by taking advantage of the most advanced radio access techniques. However, the strong dependencies between user equipment, base stations, and the Evolved Packet Core limit the flexibility, manageability, and resilience of such networks. If the communication links between UE-eNB or eNB-EPC are disrupted, mobile terminals are unable to communicate. In this article, we reshape the 4G mobile network to move toward more virtual and distributed architectures to improve disaster resilience and drastically reduce the dependency between UE, eNBs, and EPC. First, we present the flexible management entity, a distributed entity that leverages on virtualized EPC functionalities in 4G cellular systems. Second, we introduce a novel device-to-device communication scheme allowing the UE in physical proximity to communicate directly without resorting to coordination with an eNB or EPC entity.

INTRODUCTION

Fourth-generation Long Term Evolution (4G-LTE) networks are designed in such a way that the base stations (eNBs) depend on many other local and regional entities located within an Evolved Packet Core (EPC) to ensure their proper operation. This strong dependence between access and core networks limits the flexibility, manageability, and resilience of 4G-LTE systems. In fact, physical destruction of network components has been identified as the most common cause of telecommunications failures in disasters that occurred in past years. For example, during the recent Japan tsunami in 2011, a total of approximately 1.9 million fixed communication lines and 29,000 eNBs were damaged [1]. Moreover, disrupted communication links between the access and core networks affect the communications even if the network elements are working perfectly from the hardware and software point of view. Thus, problems caused by physical destruction are likely to last longer than problems caused by network congestion, for instance. Another limiting factor for resilient communications in 4G-LTE networks is the strong dependence between user equipment (UE) and the access network. In fact, UEs not only need to communicate in a traditional cellular fashion, but also need to communicate directly if the network infrastructure is temporarily unavailable or operating conditions prevent reliable communication links. In the new Release 12 of Third Generation Partnership Project (3GPP), the new device-to-device (D2D) communication feature is being specified and will help to overcome the tight interaction between UEs and the radio access network (RAN) [2].

The EPC is defined by 3GPP with the goal of providing simplified all-IP core network architecture to efficiently give access to various services. Using the user and control plane mechanisms, EPC supports a set of specialized functions to enforce access control, perform user authentication, and implement a number of application services, just to name a few. Consequently, even the breakdown of a single entity inside the EPC is capable of affecting the whole network operation. To mitigate the occurrence of cascading failures as much as possible, EPC entities employ complex techniques and equipment to provide the highest levels of reliability that are needed for mobile operators to serve hundreds or thousands of users simultaneously. Thus, the installation or replacement of hardware entities is a time consuming process that requires careful planning and the intervention of trained crews. Therefore, installing and operating such entities may cause significantly high capital (CAPEX) and operating expenditures (OPEX), with the additional drawback of reducing the resilience of the network during crisis situations.

In contrast, we focus on a software architecture design as well as a set of distributed protocols to provide higher flexibility, manageability, and resilience of 4G networks while ensuring levels of reliability similar to operational cellular network infrastructures in crisis and disaster situations. To achieve our goal of augmenting 4G network resilience, we propose a novel component within the LTE system architecture called the flexible management entity (FME), which is based on the idea of EPC entities virtualization, which entails the deployment of customized ser-

Karina Gomez, Leonardo Goratti, and Tinku Rasheed are with CREATE-NET.

Laurent Reynaud is with Orange Labs.

The research leading to these results has received partial funding from the EC Seventh Framework Programme (FP7-2011-8) under the Grant Agreement FP7-ICT-318632.

0163-6804/14/$25.00 © 2014 IEEE

vices and resource management solutions locally at the eNBs, thus reducing or eliminating the dependence on physical EPC entities. We argue that embedding the most fundamental EPC operations at the RAN side using virtual LTE-EPC entities is a fundamental step toward obtaining high-performance mobile network architectures, as discussed in [3, 4]. We also discuss a complementary solution to enable more resilient mobile networks based on an innovative D2D communication protocol where i) the necessary operations to discover, establish, and maintain D2D communications are independent of eNB and EPC entities, and ii) the LTE uplink channels (PUCCH, PRACH, and PUSCH) are used to initialize and operate the D2D protocol, making it easy to incorporate in the LTE specifications. Thus, the flexible and resilient 4G-LTE architecture model is based on an embedded eNB-EPC model that can be fully standalone and operational with the support of virtual end-to-end physical core infrastructures, and UE able to operate independent of the eNB and EPC using D2D communication mode whenever necessary.

State of the Art

EPC Virtualization

Following the idea of using virtualization for allowing resilient and shared infrastructures, the work done in [3] undertakes a discussion on alternative ways of network ownership. Similar concepts are also investigated in [5], where reconfigurable mobile network architecture is proposed for flexibility and reconfiguration. 3GPP has also recognized the importance of supporting network sharing by means of virtualization through Releases 6–12, where a set of technical specifications and architectural requirements are defined [6]. The design and implementation of a network virtualization substrate (NVS) for effective virtualization of wireless resources in cellular networks is introduced in [7]. The authors analyze the ongoing 3GPP efforts in RAN sharing and introduce a concrete implementation of an eNB virtualization solution for LTE systems. In [8], the authors introduce the distributed mobility management entity (DMME), which implements mobility management for the next generation of cellular systems using a DMME across the network.

Summarizing, [3, 5, 6] study the virtualization of components for sharing the core network, whereas [7] mainly focuses on virtualization in terms of resource sharing. On the contrary, the innovation we propose here consists of moving key EPC functions closer to the RAN to enable distributed control and management of 4G networks using virtual entities.

D2D Communication Protocols

D2D communications have recently been discussed in 3GPP in the context of the Proximity Services study item [2]. Complete D2D protocol and specifications are expected to be included in 3GPP Release 13. However, several works are already available in the existing literature. In [9] an intra-cluster D2D retransmission scheme with optimized resource utilization is presented. The scheme enhances the network throughput using the D2D cluster formation under the supervision of the eNB. Similarly, in [10] the authors propose D2D discovery and link setup procedure using eNB's supervision. A different approach is presented in [11] in which a D2D server coordinates the establishment of D2D communication links by maintaining and tracking the capabilities of the D2D UEs as well as interacting with the MME to perform D2D bearer's setup procedures. Similarly, in [12] the authors introduce network protocol and architecture for LTE-Advanced (LTE-A)-based D2D communications, where the UE plays the role of D2D enabler while the packet data network gateway (P-GW) is the D2D coordinator.

The majority of the D2D protocols available in the literature enable the D2D links with the coordination/supervision of eNB or EPC entities. We noticed that this approach weakens the resilience of the D2D network in all cases in which either the eNB or the EPC entities are unavailable, thus implying that the D2D network cannot function properly. We thereby address this issue by proposing a D2D communication scheme independent from the access and EPC entities.

Scenarios and Motivations

Crisis scenarios are often characterized by damaged network elements or severely impaired communication links between network nodes and entities. In these cases, robust architectures are required in order to keep the network services running, with low or possibly even no impact for the end users. However, the current 4G-LTE network architecture is strongly influenced by the need of monitoring the user traffic constantly. This increases the dependencies between hierarchical network entities, and greatly reduces the resilience and robustness of the whole system in case of crisis events. In 3GPP Release 11, specific restoration procedures were developed for the EPC entities. This includes MME, serving gateway (S-GW), and P-GW recovery mechanisms after a failure with and without restarting such entities [13]. However, these procedures are only detecting failures at a software level. Thus, the current 4G network is not designed for recovering from hardware failures, which is the case of the disruption caused by the occurrence of disasters that may seriously affect the provision of services to end users.

To reduce the dependence between network entities and make the network more robust, we propose a flexible 4G-LTE architecture embedding the most fundamental EPC functions inside the eNB, which we call Hybrid-eNB (HYeNB). Figure 1 illustrates an example of a flexible 4G-LTE network deployment in which the HYeNB is designed to work totally isolated from the physical EPC. Furthermore, the HYeNB supports particular functionalities of the physical EPC to enable an autonomous behavior for the provision of connectivity and services to the users or at least intranet communication for the UE. The HYeNB can rely on wired or wireless network connectivity (IEEE 802.11, 802.16, satellite, optical network, etc.) for ensuring a link to the physical EPC.

To design resilient 4G networks, we propose

To reduce the dependence between network entities and make the network more robust, we propose a flexible 4G-LTE architecture embedding the most fundamental EPC functions inside the eNB, which we call Hybrid-eNB.

Physical and virtual network entities coexist to enable distributed operations, while the D2D communication mode allows an ad hoc LTE network to be established. To allow the virtual and physical EPCs to coexist in dynamic scenarios, specific routing and topology management mechanisms were developed.

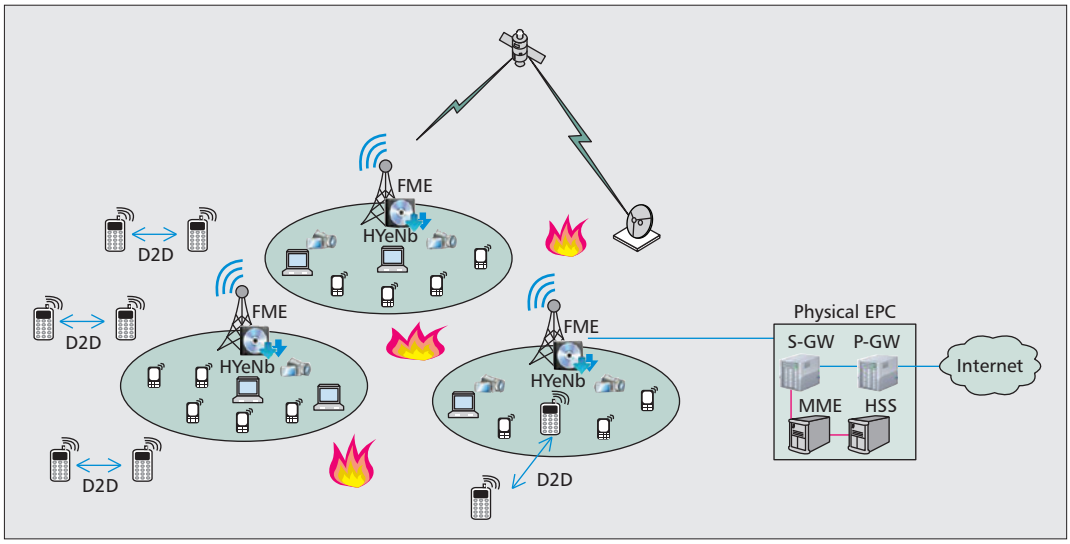


Figure 1. FME and D2D application scenarios for resilient communications.

moving from a highly centralized to a distributed system as depicted in Fig. 2. As we can observe, physical and virtual network entities coexists to enable distributed operations, while the D2D communication mode allows establishing an ad hoc LTE network. To allow the virtual and physical EPCs to coexist in dynamic scenarios, specific routing and topology management mechanisms were developed.

Flexible Management Entity Design

In this section we introduce the software elements (units and agents) and interfaces of FME (Fig. 2).

Virtual EPC

Virtual EPC (vEPC) supports specific EPC functionalities that give the HYeNB the possibility to operate autonomously by providing connectivity and other services to users. The vEPC functionalities are discussed below.

Gateway-Agent — This manages all the mechanisms implemented for supporting the basic functionalities of EPC from the user plane point of view. The gateway-agent (GW-A) is responsible for guaranteeing the proper operation of the HYeNB when it is disrupted or disconnected from the physical EPC or a physical EPC does not exist. Thus, if a specific server located in the physical EPC is temporarily unavailable (e.g., S-GW or P-GW), GW-A runs a function able to act as a surrogate server. Consequently, GW-A is responsible for providing connectivity to external packet data networks, re-establishing and performing the handover operations of all the UEs to the entities inside the physical EPC when needed.

Mobility Management Entity-Agent (MME-A) — The mechanisms implemented for enabling control plane functionalities are supported by the mobility management entity-agent (MME-A). It manages and stores UE information regarding identity, mobility state, and security parameters (Fig. 2). In fact, the MME-A interacts with the link management unit (LMU) and routing management unit (RMU), discussed in the next subsection, for supporting the creation of virtual X2 (vX2) interfaces for interconnecting HYeNBs. The MME-A is also responsible for interacting with external MME-A and RMU for performing handover procedures between vEPCs as well as periodically synchronizing the user plane contexts with the physical EPC.

D2D-Agent —The D2D-Agent (D2D-A) is responsible for managing all the mechanisms pertaining to the D2D communication mode as well as storing D2D context information. Notice that the D2D-A is meant to provide the necessary functions for allowing D2D mode communications amongst the UEs within the HYeNB coverage area only when this is required. However, UEs are also capable of creating a D2D network without relying on the D2D-A, as explained in the next section.

Link Management Unit

This unit manages the protocols of wired or wireless interfaces supported by the HYeNB in order to communicate with the physical EPC. This unit is responsible for encapsulation/de-encapsulation of all the messages exchanged between HYeNBs and the physical EPC independent of the technology supported by HYeNBs (i.e., IEEE 802.11, 802.16, satellite, etc.) This procedure actually creates a tunnel between the HYeNB and the physical EPC. Thus, LMU maintains a direct or multihop link between the HYeNBs and the physical EPC in order to i) create a virtual-S1 (vS1) interface as well as to tunnel S1 into the vS1 interface, and ii) create vX2 interfaces for interconnecting HYeNBs when the MME-A performs UE handover. The FME also supports adapted and autonomic mechanisms to avoid information loss, thus reducing the probability of service disruption during link failures or blackouts. For this

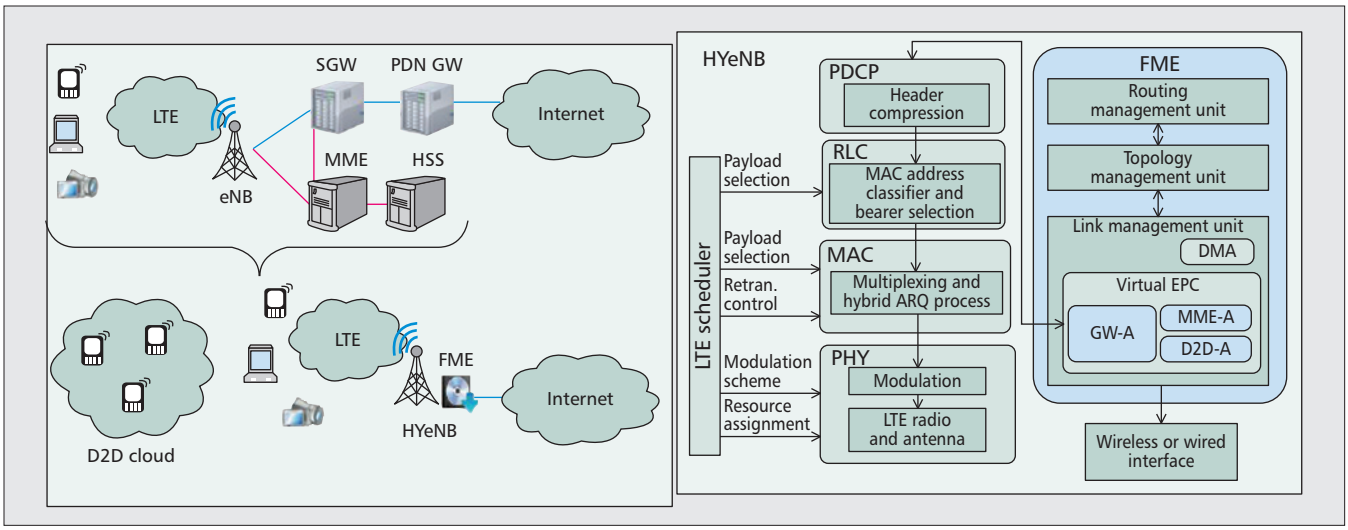


Figure 2. Resilient 4G-LTE architecture including FME software elements and interfaces.

purpose, the disruption management agent (DMA) is used.

Routing and Topology Management Units

The RMU is responsible for routing packets in the network and maintaining active paths between each HYeNB and the physical EPC. The topology management unit (TMU) is responsible for topology optimization of the HYeNBs. These units allow a dynamic topology in which the HYeNBs can always rely on a link with the physical EPC. Figure 3 summarizes the FME units and agent interactions, and the messages' handshake. More specifically, the messages exchange includes the following phases.

Creating the vX2 and vS1: When the HYeNB is activated, it first sends an interlayer discovery message to the TMU asking for information about the HYeNBs and physical EPC present in the network (message 1; the requested information includes HYeNBs and physical EPC identifiers and routes). Then the TMU sends a request message to the RMU in order to activate the routing protocols and obtain the best routes to the HYeNBs and physical EPC (message 2). Thus, the interlayer discovery and request messages are answered by the TMU and RMU in order to update the LMU with the required information (message 3). Finally, using this information, the LMU maintains a table with the HYeNBs, physical EPC ID, and the routes to create the virtual vX2 and vS1 interfaces. Moreover, the HYeNB also enables the DMA mechanism in order to support the network disconnection.

Creating the radio and Evolved Packet System bearer: When a UE requests an association with the HYeNB (message 4), the HYeNB has to interact with MME-A and GW-A for completing the request and to be able to serve the UE. Specifically, the HYeNB sends an attach request message to the MME-A (message 5). Then the MME-A and HYeNB exchange attach handshake messages (message 6), while the MME-A and GW-A exchange session request messages for completing the UE attached procedure with the HYeNB (messages 7-8-9). After that attachment and session creation procedures have been completed, the Evolved Packet System (EPS) bearer is created by the GW-A, and the UE is notified about it (messages 10-11). At this point, the UE can be served, and the intra-cell data or voice communications can be initiated.

Creating an end-to-end EPS bearer: Once the EPS bearer is established, the MME-A and GW-A transmit UE context information to the physical EPC in order to synchronize the UE information within the whole network. After that an end-to-end EPS bearer can be activated for the interaction between vEPC and EPC. At this point, the inter-cell data or voice communications can start.

The FME units (LMU, TMU, and RMU) interact with one another to create and maintain the vX2 and vS1 interfaces, while the FME agents (GW-A and MME-A) create and maintain the bearers of the different services required for serving the UEs (Fig. 3). Consequently, UEs already associated can be served immediately for inter-/intra-cell communications. In disaster scenarios, the services should not follow usual billing rules because:

- Billing operations use available resources that are needed for other purposes in disasters and emergencies.
- Billing operations would prevent the communications of users without credit, which can be dangerous in life-threatening conditions.

To conclude, the FME is a software solution that distributes several core functionalities at the access side whereby software units and agents allowing the eNB to keep the network services running with low impact for the end user in case of network disruption.

D2D Communication Scheme

While FME reduces the dependence between access and core networks, mechanisms for allowing UEs to operate independent of the access network are also required. In this way, 4G networks (but, even more important, future 5G communications systems) will leverage resilient

The main advantage of using SC-FDMA in D2D communications is to protect cellular downlink communications from interference and to spare battery. We also make the assumption that frequency-division duplexing is used.

communications similar to a tactical network. For this reason, we propose a D2D communication protocol where the UE itself is the enabler, coordinator, and manager of the D2D network even without any preliminary interaction with the access network. Thus, when connectivity with the eNB is lost or nonexistent for at least a time of interruption (ToI), a selected UE is responsible for establishing and managing the D2D network using single-carrier frequency-division multiple access (SC-FDMA) as modulation format for transmission and reception. The adoption of a ToI is necessary to avoid undesired ping-pong effects. The main advantage of using SC-FDMA is to protect cellular downlink communications from interference and to spare battery. We also make the assumption that frequency-division duplexing (FDD) is used. The salient features of the proposed D2D communication protocol are presented below.

D2D Network Discovery

If the HYeNB is not available for at least a ToI, a UE is allowed to start the procedure for creating or joining a D2D network. Thus, a UE, which is denoted as *b-UE*, evaluates the possibility of broadcasting direct beacon frames (D-beacons). We refer to a D-beacon interval (TD) as the period between two consecutive beacon transmissions. For commercial users, in principle any UE could transmit a beacon frame, but before doing that UEs must listen to the channel for at least two D-beacon intervals in order to check whether a D2D network already exists, and avoid collisions and interference.

For public safety users, in general only the UE of the chief of a public safety group can transmit a beacon frame to create a private and secure D2D network. Thus, both features of the D2D protocol will rule the creation of D2D networks in the beginning. The D-beacon interval is supposed to be any multiple integer of the LTE radio frame duration (i.e., 10 ms). If a beacon is not received, the UE is allowed to start broadcasting its own beacon frame. The transmission of D-beacons by the b-UE has the purpose of replacing the signals from the HYeNB for other UEs. Some of the information periodically conveyed by the D-beacon frame is, for example, the D2D network identification and the identity of all the UEs connected to that particular network. In order to exploit the available uplink resources and structure, the physical uplink control channel

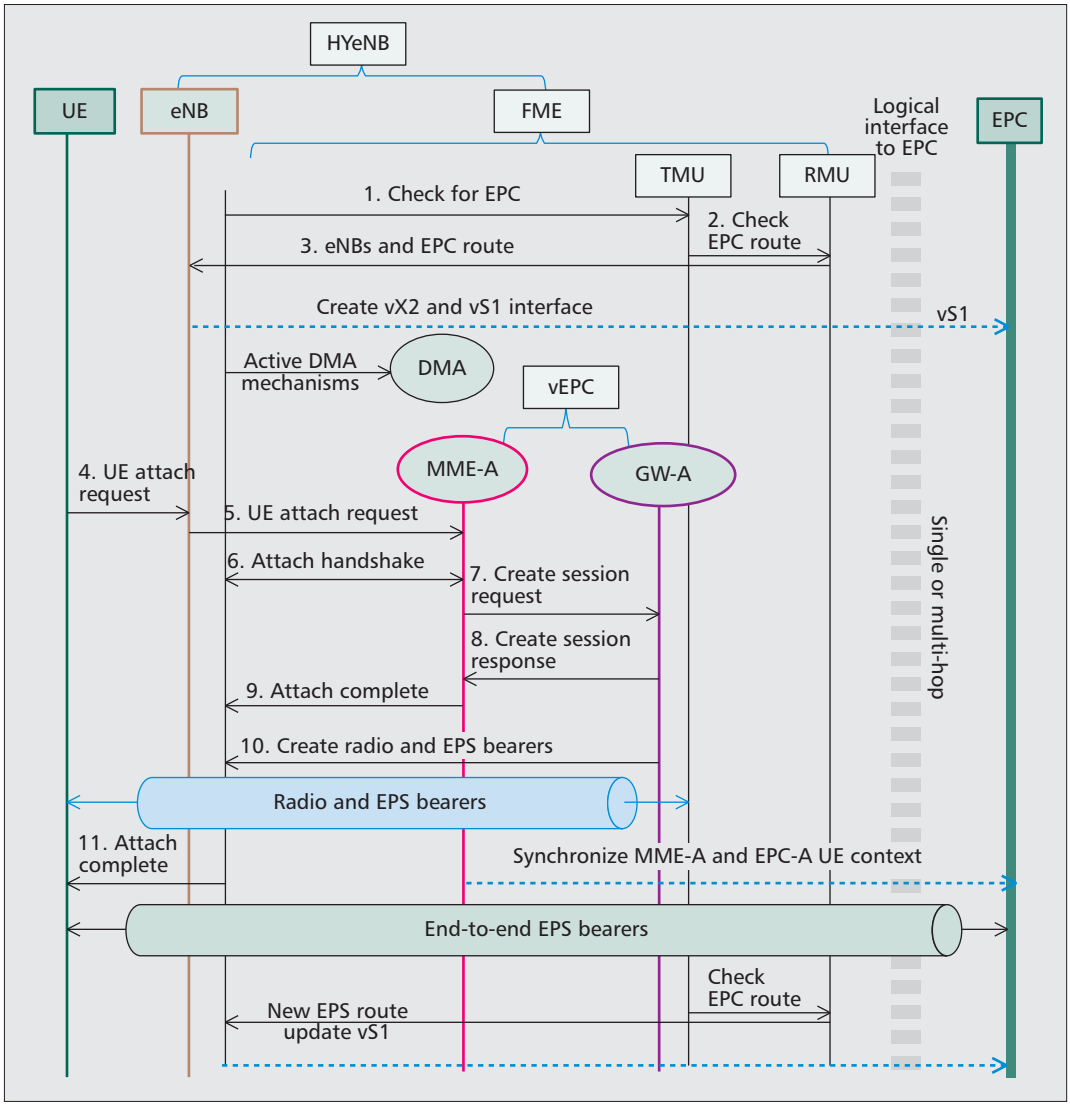


Figure 3. FME handshake message overview.

(PUCCH) is exclusively used by the b-UE to transmit D-beacons (Fig. 4) and to reply to network association requests of other UEs.

Notice that in case the HYeNB is available, the authorization and initialization of D2D mode communications can be controlled by the D2D-A, which is available in the FME architecture shown in Fig. 2. Consequently, for the purposes of coverage area extension and traffic offloading from the HYeNB, the FME allows UEs inside the coverage area to setup D2D communication sessions.

D2D Network Establishment

After receiving D-beacons broadcast by the b-UE, the UEs can join the D2D network. Thus, UEs trying to establish a connection with the b-UE shall follow the four-way handshake defined in standard LTE specifications and that define random access channel (RACH) operations as shown in Fig. 4. Logical RACH functions are carried over the physical RACH (PRACH) slots that are available in fixed positions inside the structure of the D-beacon interval as defined by the b-UE. We assume that the PUCCH is also used by the b-UE to respond to association/authentication requests of other UEs. To join the D2D network, as in the normal LTE procedure, UEs randomly select one of 64 Zadoff-Chu (ZC) preamble signatures, thus mitigating the risk of potential collisions. In fact, we assume that multiple independent ZC sequences can be correctly decoded by the b-UE. We propose that the response of the b-UE shall follow over the PUCCH rather than over the physical downlink shared channel (PDSCH) as in standard LTE but still in contentionless mode. Furthermore, we also allocate reserved slots in the physical uplink shared channel (PUSCH) to carry the third message sent by the UEs during the four-way handshake, as well as to reserve resources for the purpose of exchanging data or setting up voice calls (additional PRACH slots). Notice that reservation of resources for data and voice does not involve the b-UE but rather takes place among peer entities in a distributed manner. The preamble contention resolution shall follow the standard LTE procedure (i.e., the b-UE assigns D2D network identifications). At this stage we also envisage room to apply different random backoff schemes in order to reduce collisions over the same preamble sequence. Data communications will then take place over the reserved slots within the PUSCH.

D2D Network Disassociation

In order to enable timely discovery of the HYeNB, we suppose that periodically all UEs (including the beaconing terminal) shall make attempts to detect synchronization signals and the master information block (MIB) sent by the HYeNB over the physical broadcast channel (PBCH). Indeed, we assume that if a particular UE manages to recover connectivity with the HYeNB, resources allocated to that D2D network must be relinquished. In this way FME objectives are complemented by enabling the UEs with the capability of forming ad hoc topologies, thus drastically reducing the dependence between user equipment, access, and core network elements.

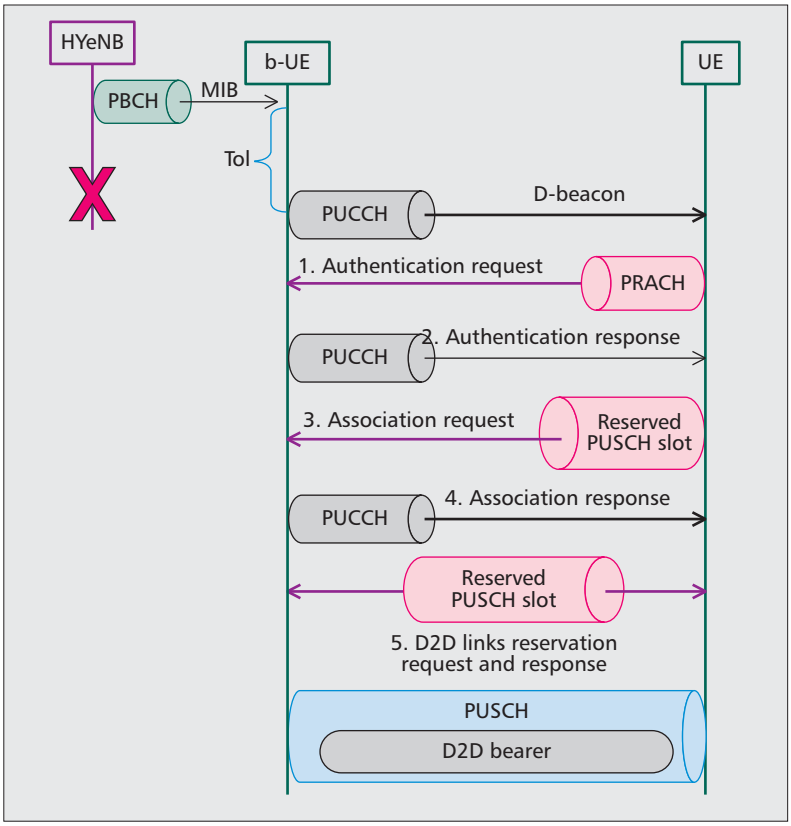


Figure 4. Handshake messages for joining a D2D network.

Performance Analysis

We present here the evaluation of the advantages arising from the use of FME and D2D communication protocol for resilient 4G networks. A network suitable for public safety scenarios (Fig. 1) was simulated in an LTE module in the OMNeT++ simulator developed by the authors [14]. We consider a scenario over a surface of 2 km^2, and we assume that a natural disaster cuts off communication links between the HYeNB and the EPC. In the disaster area, three HYeNBs are deployed with a distance of approximately 1 km between each other. We assume that a physical EPC is located approximately 1.5 km away from the disaster area. Furthermore, only a single HYeNB has direct connectivity with the physical EPC. The HYeNB backhaul network is created using an ad hoc WiFi link. It is worth noticing that other technologies regarding the backhaul network have been examined by the authors. However, the main contribution of this work focuses on validating the role of FME in managing dynamic scenarios. For the sake of D2D networks, we assume that on average $M = 75$ UEs (realistic value for the preliminary phase in rescuing operations) are outside the coverage area of the HYeNBs, and thus they are allowed to set up D2D mode communications. Table 1 summarizes the parameters used for modeling LTE, WiFi, and D2D link parameters. For FME validation simulations, the cells are configured in time-division duplexing (TDD) mode with 10 MHz bandwidth using 50 full resource allocations, 16-quadrature amplitude modulation (QAM), and slot configuration num-

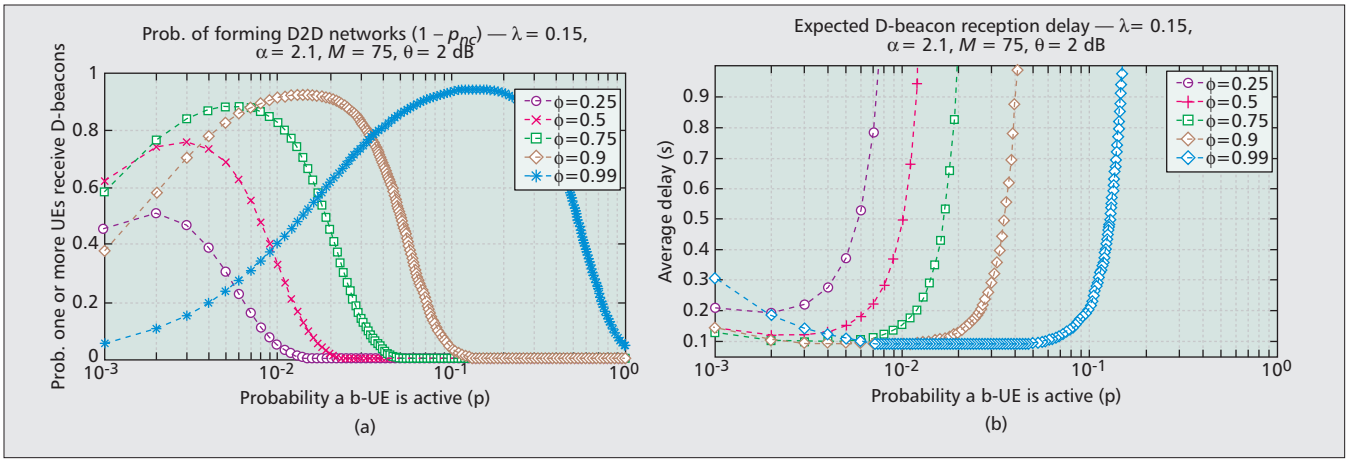


Figure 5. D2D network connectivity probability and average access delay assuming a total of $M = 75$ UEs: a) probability that at least a D-beacon is received correctly by UEs; b) D-beacon reception delay while varying the probability that a b-UE is active.

ber 1, where the slots dedicated for DL/UL are the same. Inside each cell, 250 UEs are uniformly distributed. The UEs move inside the coverage area following a random waypoint mobility model with a uniformly distributed speed between (0.2–0.7) m/s. The following multimedia applications, with values realistic for the initial phase in rescuing operations are simulated:

- 40 intra-cell calls and 5 UL/DL real-time video streaming between UEs and the video server for each cell. The video server is located inside the physical EPC.
- 20 inter-cell calls between cell-1/cell-2 and cell-2/cell-3.

The inter-cell calls were simulated using an enhanced voice services (EVS) codec encoded at 64 kb/s, and real-time video streaming was simulated using an H.264 codec encoded at 384 kb/s. The applications run in parallel for 600 s. Results presented in this section were averaged over 10 simulation rounds. For D2D, a path loss exponent equal to 2.1 was used (similar to free-space propagation). We assume a D-beacon interval TD of 80 ms and a D-beacon duration equal to a 10 percent overhead of the whole TD.

With regard to FME simulation results, the main findings are discussed below:

- Using FME, the HYeNBs work properly even if they are disconnected from the physical EPC. Since the MME-A and GW-A are able to substitute physical EPC functionalities, the inter-/intra-cell calls can be established using their capabilities. Notice that these scenarios are not possible using the standard LTE architecture.
- The achieved throughput in UL and DL where similar (around 10 Mb/s) as both use similar configuration distinguished only by the amount of overhead introduced. While the average UL/DL throughput achieved for each served user is similar for cell 1 and cell 3 (around 140 kb/s in UL and 170 kb/s in DL), the average UL/DL throughput for cell 2 is less than that of cell 1/cell 3, since cell 2 has more users (around 110 kb/s in UL and 130 kb/s in DL).
- We conclude by postulating that the vEPC is able to perform routing functionalities for serving the cells and allowing inter-/intra-cell calls in the HYeNB networks. Furthermore, it allows video streaming from external networks passing through the physical EPC.

In the case of D2D, Fig. 5a shows the probability of forming D2D networks in non-coverage regions. The results are obtained assuming that UEs are scattered over the 2D Euclidean space according to a homogeneous Poisson point process of intensity λ_s, as explained in detail in [15]. The parameter Φ in the figure denotes the fraction of regular UEs with respect to the total (i.e., non-beaconing UEs). Essentially the analysis investigates the connectivity aspect of the D2D network and the delay incurred in the D2D network formation. As depicted in the figures, this allows identifying optimal values of the probability (p) a b-UE is actively sending D-beacon frames. The optimum value of probability tends to shift from left to right as the density of regular UEs is increased (vice versa, the density of b-UEs is decreased), and more than a single value the curves show almost a narrow region. Decreasing the number of active b-UEs implies that we need more of them transmitting beacons in order to ascertain connectivity inside the region. Based on these results, we also obtain the D-beacon reception delay shown in Fig. 5b, which allows quantifying on average the time that UEs out of coverage might be completely disconnected, which is the critical time lag between the instant UEs lose connectivity with the HYeNB and the instant UEs join the D2D network. Consequently, the proposed D2D protocol shows the capability of improving the communication resilience of UEs outside network coverage.

Conclusions

We present FME, a novel architectural solution to realize simplified and virtualized EPC functions, which is the enabler for deploying and managing resilient eNBs that are capable of standalone and autonomous operations. This reduces CAPEX and OPEX, as well as time and effort in redeploying a multi-service, multi-band, interoperable, and integrated network infra-

structure for specific applications, including emergency communications. FME is an example of a software-defined solution for running the vEPC inside the edge for the provisioning of services even when the physical EPC fails or is unavailable. Aiming to achieve more standalone network operations, we also propose a novel D2D communication protocol endowing LTE-enabled mobile user equipment with ad hoc capabilities. The advantage of the proposed protocol is the flexibility of starting D2D mode communications without any interaction with the radio access network.

References

[1] "ICT responses to the Great East Japan Earthquake, FUJINO, Masaru. Counselor for Communications Policy. Embassy of Japan," U.S. Telecom Assn. Boarding Room, Dec. 6, 2011.
[2] 3GPP TDoc S1-120349 (Draft of TR 22.083), "Feasibility Study for Proximity Services (ProSe)", Feb 2012, http://www.3gpp.org/Releases.
[3] T. Forde, I. Macaluso, and L. Doyle, "Exclusive Sharing and Virtualization of the Cellular Network," *Proc. IEEE DySPAN*, 2011, pp. 337–48.
[4] "Stateless User-Plane Architecture for Virtualized EPC," http://tools.ietf.org/html/draft-matsushima-stateless-uplane-vepc-01.
[5] A. Khan *et al.*, "The Reconfigurable Mobile Network," *Proc. IEEE ICC*, 2011, pp. 1–5.
[6] "3GPP TS 22.951," Service Aspects and Requirements for Network Sharing," v. 11.0.0, 2012, http://www.3gpp.org/Releases.
[7] X. Costa-Perez *et al.*, "Radio Access Network Virtualization for Future Mobile Carrier Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, 2013.
[8] X. An *et al.*, "dMME: Virtualizing LTE Mobility Management," *IEEE Conf. Local Computer Networks*, 2011, pp. 528–36.
[9] B. Zhou *et al.*, "Intracluster Device-to- Device Relay Algorithm with Optimal Resource Utilization," *IEEE Trans. Vehic. Tech.*, vol. 62, no. 5, 2013, pp. 2315–26.
[10] J. Hong *et al.*, "Analysis of Device-to-Device Discovery and Link Setup in LTE Networks," *Proc. IEEE PIMRC*, Sept. 2013, pp. 2856–60.
[11] B. Raghothaman *et al.*, "Architecture and Protocols for LTE-Based Device to Device Communication," *Proc. ICNC*, 2013, pp. 895–99.
[12] M. J. Yang *et al.*, "Solving the Data Overload: Device-to-Device Bearer Control Architecture for Cellular Data Offloading," *IEEE Vehic. Tech. Mag.*, vol. 8, no. 1, 2013, pp. 31–39.
[13] 3GPP TR 23.857, "Study of Evolved Packet Core (EPC) Nodes Restoration," http://www.3gpp.org/Releases.
[14] K. Gomez *et al.*, "Performance Evaluation of Broadband Aerial LTE Base-Stations for Emergency Recovery," *Proc. GLOBECOM 2013*, Dec. 2013.
[15] L. Goratti *et al.*, "A Novel Device-to-Device Communication Protocol for Public Safety Applications," *Proc. IEEE D2D Wksp. at GLOBECOM*, Dec. 2013.

Link	LTE UL/DL	D2D	WiFi UL/DL
Channel model	Clarke's Fading 20 fading paths	Free Space	Free Space
Modulation	16-QAM 1/3(SISO)	QPSK	BPSK/QPSK/QAM
α	2.2	2.1	2
Transmission power	UE: 23 dBm eNB: 30 dBm	23 dBm	23 dBm
Central frequencies	700 MHz	700 MHz	5 GHz
Antenna gain	3/3 dB	0/0 dB	3/3 dB
Receiver sensitivity	UE: –107.5 dBm eNB: –123.4 dBm	–107.5 dBm	–85 dBm
Thermal noise	UE: –104.5 dBm eNB: –118.4 dBm	–174 dBm/Hz	–110 dBm

Table 1. Simulation parameters for channel propagation.

Biographies

Karina Mabell Gomez Chavez (karina.gomez@create-net.org) received her Master's degree in wireless systems and related technologies from the Turin Polytechnic, Italy, in 2007. In 2007, she joined FIAT Research Center, becoming part of Infomobility-Communication and Location Technologies. In July 2008, she joined the iNSPIRE Area at CREATE-NET, working on several projects She received her Ph.D. degree in telecommunications from the University of Trento, Italy, during 2013. During her Ph.D. she conducted various industry internships and research visits at Orange Labs, Lannion, France, Telekom Innovation Laboratories, Berlin, Germany, and the School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia. Since the beginning of 2013, she has been part of the FuN Area at Create-Net. She has several patents on protocols for next generation wireless networks and has published her research in important journals and conferences.

Leonardo Goratti (leonardo.goratti@create-net.org) received his Ph.D. degree in wireless communications in 2011 from the University of Oulu, Finland, and his M.Sc. in telecommunications engineering in 2002 from the University of Firenze, Italy. From 2003 to 2010, he worked at the Centre for Wireless Communications (CWC), Oulu, Finland, first as a researcher and then as a Ph.D. student. His research interests cover MAC protocols for wireless personal/body area networks and wireless sensor networks, as well as routing protocols for sensor networks. His research interests also cover UWB transmission technology and 60 GHz communications. From 2010 until early 2013 he worked on MAC protocols for cognitive radios and spectrum sharing techniques at the European funded Joint Research Centre (JRC) of Ispra, Italy. Recently he joined the Research Centre CREATE-NET, Trento, Italy, where he is currently working on the European project ABSOLUTE on LTE-based device-to-device communications in the context of public safety scenarios.

Tinku Rasheed (tinku.rasheed@create-net.org) is a senior research staff member at Create-Net. Since May 2013, he is heading the Future Networks R&D Area [FuN] within CREATE-NET. Before joining CREATE-NET in December 2006, he was a research engineer with Orange Labs R&D from May 2003 to November 2006. He received his Ph.D. degree from the Computer Science Department of the University of Paris-Sud XI in 2007. He completed his M.S. degree in 2003 from Aston University, United Kingdom, specializing in telecommunication engineering and his Bachelor's degree in 2002 in electronics engineering from the University of Kerala, India. He has extensive industrial and academic research experience in the areas of mobile wireless communication and data technologies, and end-to-end network architectures and services. He has several granted patents and has published his research in major journals and conferences.

Laurent Reynaud (laurent.reynaud@orange.com) is a senior research engineer and expert for the Future Networks research community at Orange. Specializing in the performance of agile infrastructures for challenging environments, his topics of interest include wireless multihop, multi-path, and large-scale routing techniques, as well as various issues related to QoS in wireless networks. After receiving his engineering degree from ESIGETEL at Fontainebleau in 1996, he acquired significant experience regarding the development and deployment of distributed software in the context of telecommunications through successive positions in the French Home Department in 1997, Alcatel-Lucent from 1998 to 2000, and at Orange since 2000. He has participated in multiple French, European, and international cooperative research projects. He has co-authored many conference and journal articles, and holds a dozen international patents.

EmergeNet: Robust, Rapidly Deployable Cellular Networks

Daniel Iland and Elizabeth M. Belding

ABSTRACT

Cellular phone networks are often paralyzed after a disaster, as damage to fixed infrastructure, loss of power, and increased demand degrade coverage and quality of service. To ensure disaster victims and first responders have access to reliable local and global communication, we propose EmergeNet, a portable, rapidly deployable, small-scale cellular network. In this article, we describe EmergeNet, which addresses the challenges of emergency and disaster areas. EmergeNet provides free voice calling and text messaging within a disaster area, and enables users of unmodified GSM handsets to communicate with the outside world using the Skype VoIP network. We evaluate EmergeNet's ability to provide robust service despite high load, limited bandwidth, and software or hardware failures. EmergeNet is uniquely well suited to providing reliable, fairly allocated voice and text communication in emergency and disaster scenarios.

INTRODUCTION

Cellular telephony is the most widely adopted communication technology on the planet. There are more than 6.8 billion active cellular subscriptions worldwide, with half of those subscriptions added in the last five years, largely in developing areas [1]. In many developing areas, cellular networks have leapfrogged traditional landline telephone and Internet infrastructure due to the comparative ease and low cost of deployment.

Investments in cellular network infrastructure are often concentrated in urban areas and along transport corridors. The combination of low population density and lack of reliable infrastructure makes commercial-grade cellular network deployments largely unprofitable in rural areas. As a result, low-density areas, rural areas, and areas with limited power infrastructure are largely underserved by cellular infrastructure.

Community cellular networks have emerged to address this coverage gap, offering economically sustainable cellular service in rural areas using low-cost cellular base stations. Recent deployments demonstrate the technological and economic feasibility of such networks. For example, the University of California (UC) Berkeley researchers operate a profitable cellular deployment in Papua [2]. Rhizomatica, a community-owned non-profit, operates five GSM networks in the Mexican state of Oaxaca, providing unlimited local calling and texting to residents for $1.20 per month [3].

Cellular base stations deployed by community cellular networks can cost under $10,000, while a large cellular carrier might spend $100,000 or more to deploy a base station. A complete cellular base station can be constructed with only a few components: a Linux PC, a software defined radio, a power source, an amplifier, and an antenna. Open source software running on each base station, such as OpenBTS, handles Um interface communication with GSM devices, and translates GSM calls and text messages to Session Initiation Protocol (SIP) messages. Telephony software, such as FreeSWITCH, enables users to make and receive calls, send and receive text messages, and access interactive services through voice calls or text messages. Community cellular networks use IP infrastructure, such as Wi-Fi networks and the Internet, for communication between base stations and with the rest of the world.

As it turns out, many of the characteristics of a successful community cellular network are also essential for emergency and disaster networks. We define emergency and disaster networks (EDNs) as communication networks used by first responders and victims of disasters. Both types of networks operate under challenging power and network backhaul conditions. Both must be easy to deploy, operate, and maintain. EDNs have the additional requirements of serving large numbers of users, dealing with attempted usage in excess of network capacity, and operating in rapidly changing networking environments.

EDNs are used to provide local communication inside a disaster area and enable contact with the rest of the world. These networks may use pre-existing wireless infrastructure, or first responders may build and deploy a network as part of their recovery efforts. Often, it will be a combination of both, as all available avenues for connectivity are explored. EDNs are rapidly evolving networks, as additional hardware is added, backbone infrastructure is restored, and

The authors are with the University of California at Santa Barbara.

0163-6804/14/$25.00 © 2014 IEEE

users migrate from network to network. One example of an EDN, the Red Hook Wi-Fi network in Brooklyn, New York, went from serving dozens of users to more than 300 users per day after Hurricane Sandy impacted the area [4]. The performance of the network was impacted as usage exceeded network capacity. As the network's importance as a community information hub grew, the network was expanded to include additional Wi-Fi access points and a satellite backhaul connection from the Information Technology Disaster Resource Centre [5].

Emergency and disaster networks that incorporate GSM cellular base stations have several advantages over non-cellular EDNs. A single cellular base station can cover a radius of up to 3 km, providing a coverage area equivalent to dozens of Wi-Fi access points [6]. GSM handsets are ubiquitous worldwide, and while Wi-Fi devices are very popular, Wi-Fi does not yet have the global market penetration of GSM, particularly in the developing world. This makes cellular-based EDNs an ideal means of providing communication services to a large percentage of the world population, and providing wireless coverage to large areas without extensive hardware deployments.

Lack of reliable power infrastructure, a chronic problem in developing rural areas, can paralyze cellular infrastructure in even the most well developed areas after a disaster. Following Hurricane Sandy, the FCC reported approximately 25 percent of cell sites in the affected 10-state area were non-operational [7]. Commercial cellular base stations require access to network operator services such as a mobile switching center and home location register. Damage to central switching stations, power outages, or loss of backhaul connections can eliminate cellular service across large areas. Community cellular base stations do not require any remote infrastructure to operate, and are therefore better suited to emergency and disaster use.

In this work, we leverage a rural community cellular network, VillageCell (also called Kwiizya), as a starting point for a rapidly deployable cellular system for emergency and disaster networks [6, 8]. Importantly, we augment VillageCell to make it suitable for EDNs by introducing features to enable inbound and outbound calling and messaging, handle high load, and enable automatic reconfiguration to address changing power and networking scenarios. We call this new solution EmergeNet. Our core contributions in this work are:

- Providing inbound and outbound VoIP calling and messaging to EDN users with unmodified GSM handsets
- Designing an SMS-based call queuing system to provide fair access to voice services during periods of high load
- Enabling automatic reconfiguration of cellular base stations to maximize functionality in the face of power, network, and hardware failures

We begin by describing VillageCell, the community cellular network on which EmergeNet is based. We then discuss the limitations of VillageCell base stations as components of an EDN, which inform the design of EmergeNet. We address these limitations with the design and implementation of EmergeNet, a community cellular network for emergency and disaster use. Finally, we evaluate the performance and effectiveness of EmergeNet as a means of communication in EDNs.

Introduction to VillageCell

VillageCell is a low-cost community cellular network designed for rural areas based on open source software and low-cost hardware. VillageCell was designed to allow local users to communicate by voice or SMS with other local users for free. A VillageCell deployment, such as our trial deployment in Macha, Zambia, consists of one or more cellular base stations linked together using a Wi-Fi network [8]. Our previous work revealed that technological communication in rural areas most often occurs between users who live in the same village [9, 10]. By using several cellular base stations to provide GSM cellular coverage to large areas of a village, VillageCell enables local users to communicate with each other.

Operation and Design

Each cellular base station in a VillageCell network consists of a Linux PC with a software defined radio running OpenBTS. Additionally, one VillageCell base station in each network must operate three required network services: sipauthserve for user management and access control, smqueue for queuing and delivering text messages, and the FreeSWITCH PBX for routing calls and text messages. Figure 1 shows the interactions between each of these software components in a VillageCell network. Figure 1 also shows EmergeNet's integration with the Skype network, which is described in the next section.

OpenBTS is an open source program that uses a software-defined radio transceiver to communicate with standard GSM mobile phones. OpenBTS can operate in the 850, 900, 1800, or 1900 MHz range. OpenBTS handles all communication with GSM handsets, which occurs on the GSM Um interface. OpenBTS converts GSM messages into equivalent SIP messages, and sends SIP messages (e.g., registration, call initiation, and text messages) to sipauthserve or FreeSWITCH. OpenBTS listens for incoming SIP messages, and converts them to GSM transmissions on the Um interface. OpenBTS also encapsulates GSM voice calls into Real-Time Transport Protocol (RTP) audio streams, suitable for switching by FreeSWITCH. The design of OpenBTS ensures that FreeSWITCH, smqueue, and sipauthserve do not need to communicate over the Um interface to interact with GSM phones. Instead, they send SIP messages to OpenBTS. This greatly reduces the complexity of communicating with GSM devices.

When a user enters the coverage area of the VillageCell network, her phone may connect automatically to a VillageCell base station. Most GSM handsets will roam on any available GSM network if their home network is not available. If commercial cellular service is available, the

Lack of reliable power infrastructure, a chronic problem in developing rural areas, can paralyze cellular infrastructure in even the most well developed areas after a disaster. Following Hurricane Sandy, the FCC reported approximately 25 percent of cell sites in the affected 10-state area were non-operational.

user will stay on her home network unless she manually selects the EmergeNet network in her phone's settings. As GSM handsets perform no authentication of GSM base stations, the EmergeNet base station may provide service to all mobile users, without any a priori information from cellular operators about their customers.

When a VillageCell user calls another user, OpenBTS receives GSM call establishment messages from the phone, converts them into a SIP message, and sends it to FreeSWITCH. FreeSWITCH routes call requests and bridges the call's RTP audio stream to the recipient's OpenBTS base station. When a VillageCell user sends a text message, the SMS message is received by OpenBTS and sent to FreeSWITCH for processing. This allows SMS applications to be implemented as scripts called by FreeSWITCH when a message is received. To deliver text messages to GSM handsets, FreeSWITCH sends the text message to smqueue, which makes a best effort delivery to the OpenBTS base station where the recipient is registered.

Limitations of VillageCell

While VillageCell is ideally suited for rural environments, it has a number of important limitations in the context of EDNs. These include the existence of single points of failure, capacity limitations, and a lack of connectivity to the rest of the world, which we describe in more detail below.

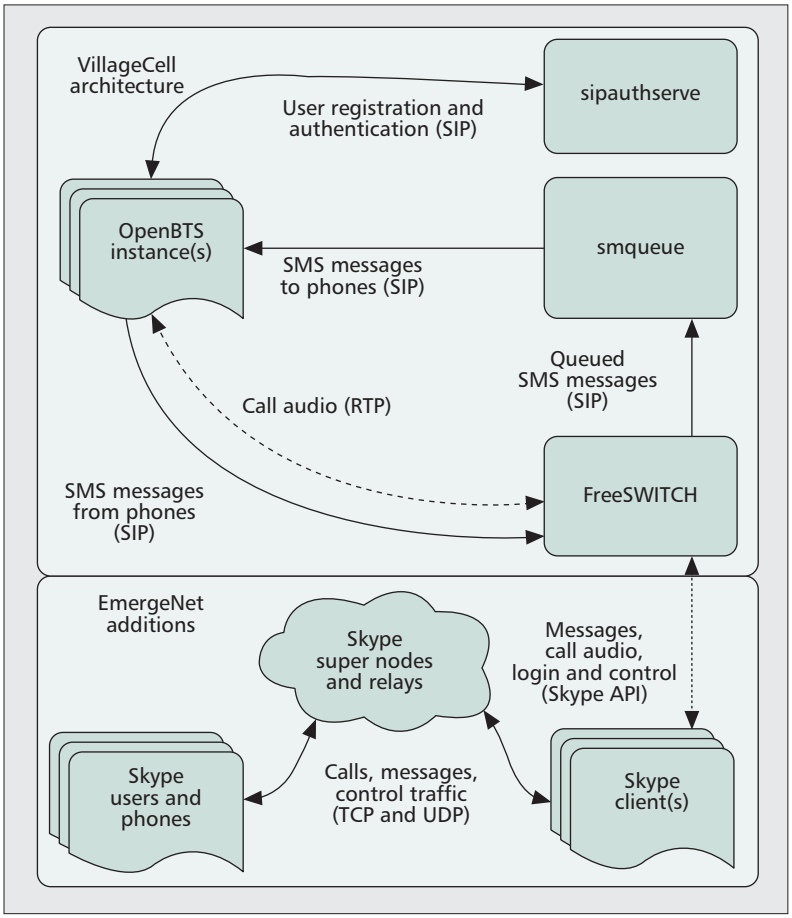


Figure 1. Architecture of VillageCell and EmergeNet.

Single Points of Failure — The vulnerability of VillageCell to single points of failure is exemplified in the specific trial deployment in Macha, Zambia. In this deployment, the base station hosting FreeSWITCH, sipauthserve, and smqueue was struck by lightning. The base station's network connection was rendered nonfunctional. The remaining functional base stations did not receive responses to the SIP packets they were generating as users attempted to place calls and send messages. Because FreeSWITCH did not respond, calls were not connected and messages were not sent, even when both users were connected to the same OpenBTS base station. Any network failure impacting the main base station, or failure of the main base station itself, rendered the entire network unusable.

Capacity Limitations — Each of the cellular base stations deployed in Macha is a Range Networks Snap Unit, consisting of a Linux PC and a RAD1 software defined radio with one absolute radio frequency channel number (ARFCN). Each base station can support about 15 registrations/min, 40 text messages/min, and 7 concurrent calls [11]. When a base station reaches the maximum number of concurrent calls, further call attempts result in calls being rejected due to congestion. VillageCell does not make any special considerations for operation under particularly high demand, which can negatively impact the user experience.

Inbound and Outbound Communication — VillageCell was designed to facilitate local communication. As such, it does not include support for calling or texting out of the VillageCell network. In a post-disaster environment, communication with the outside world is critical. Disaster victims and first responders will seek assistance from outside the impacted area by contacting aid groups, friends and family, insurance agents, response coordinators, and other remote people or groups.

EmergeNet

We use our experience with VillageCell to introduce EmergeNet, a community cellular network architecture designed for robust performance in EDNs. We augment VillageCell to make it suitable for EDNs, by introducing features to enable inbound and outbound calling and messaging, handle periods of high load, provide for rapid deployment, and enable automatic reconfiguration to mitigate the impact of power, network, and hardware problems. The following sections explain each of these features in more detail.

Inbound and Outbound Communication

While VillageCell focused on providing free local calling and texting, EmergeNet addresses the need for bidirectional communication with the rest of the world. We enable "inbound" calling and text messaging, which originates from any phone or Skype user in the world and terminates at a cellular phone connected to an EmergeNet base station. We also allow for "outbound" calling and text messaging, where

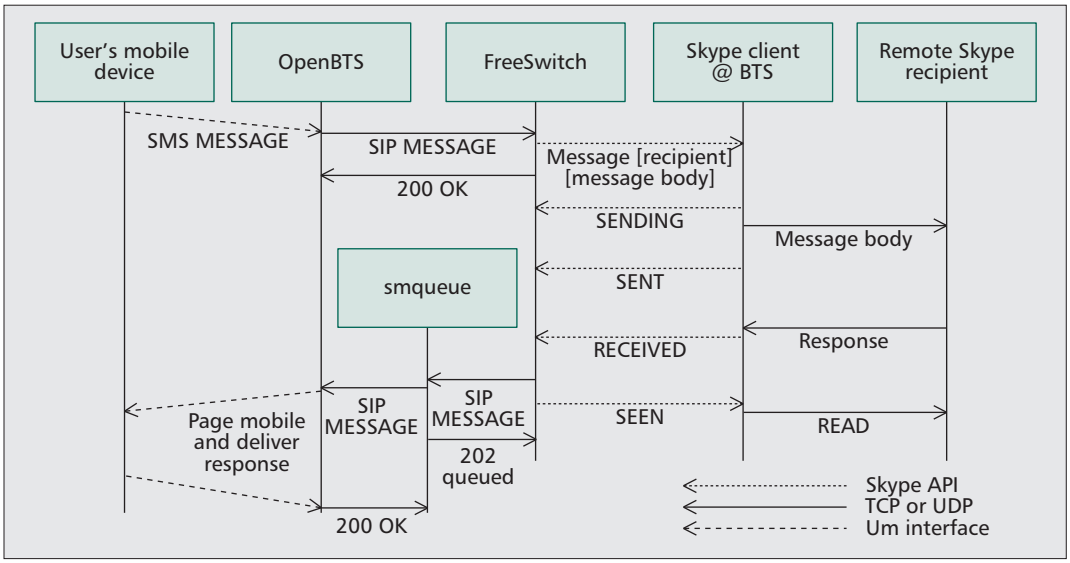


Figure 2. Messaging between a GSM phone user and a Skype user.

While VillageCell is ideally suited for rural environments, it has a number of important limitations in the context of EDNs. These include the existence of single points of failure, capacity limitations, and a lack of connectivity to the rest of the world.

communication originates in the EmergeNet network and terminates at any Skype user or phones in 170 countries. We have developed VillageVoIP, a set of software tools that runs on a cellular base station, and enables inbound and outbound calling and messaging through Skype.

VillageVoIP allows anyone with a GSM handset to use Skype, which usually requires an Internet connection and a smartphone or PC. We do not require end users to possess any software on their device. Instead, we run a Skype client for each user on an EmergeNet base station. To minimize resource consumption, each Skype client is launched using a virtual X window session, and a "fake" sound driver. Users control their Skype session by sending SMS messages to the base station.

The SMS interface to Skype has five commands:

1 login [username] [password]: Starts a Skype session for the specified Skype account
2 logout: Terminates the user's Skype session
3 friends: Sends the user an SMS listing their online Skype contacts
4 chat [recipient] [message]: Sends a text message to the recipient
5 call [recipient]: Initiates a Skype call to the recipient

VillageVoIP uses scripts launched by FreeSWITCH to interact with users via SMS or voice calls. Figure 2 shows a round-trip SMS message flow between a user's phone and a remote Skype user. The user's outbound SMS message is passed from OpenBTS to FreeSWITCH, then parsed by a python script. The python script uses the FreeSWITCH endpoint mod_skypopen to pass Skype application programming interface (API) commands to Skype clients from within FreeSWITCH. When inbound Skype communication is received, python scripts within FreeSWITCH pass chat messages to smqueue for delivery to users via SMS, or trigger a voice call by sending SIP messages to OpenBTS. If a user logs into their own Skype account, inbound chat messages and Skype calls will automatically be routed to their cellular device. If a user does not log into their own Skype account, their outgoing chat messages and calls originate from a shared Skype account.

Calling any Skype user or toll-free numbers in many countries, including Australia, France, Germany, the United States, and Taiwan, is free. This will enable an EmergeNet user to contact friends, family, coworkers, and nonprofit organizations without charge. Usage of non-free Skype features, such as calling landline phones, will be automatically billed to the logged-in user's Skype account.

Handling Periods of High Load

Previous experiments have shown that as SMS message load increases on OpenBTS platforms, delay in sending or receiving SMS does not substantially increase [6]. SMS messages are asynchronous, and typically are automatically retried by the user's handset if not immediately sent. Therefore, queuing or congestion are unlikely to negatively impact a user's SMS experience.

In an emergency situation, it is nearly guaranteed that attempted use of each EmergeNet base station will exceed seven concurrent calls, the maximum call capacity of EmergeNet's OpenBTS system. Once this limit is reached, further call attempts will result in the calls terminating seconds after dialing. Thus, the limit on concurrent calls, if not addressed, will lead to a negative user experience in an already stressful situation. It will also lead to an increase in congestion as users repeatedly retry their calls.

We mitigate this issue by creating a queue for voice calls. If demand for voice calls exceeds the base station's voice call capacity, each additional call request is placed in a first-in first-out queue, and the caller is informed of their place in the queue via SMS. A script monitors channel availability and CHANNEL_HANGUP_COMPLETE events generated by FreeSWITCH, and connects calls in the queue as channels become available. This prevents users from having to repeatedly retry their calls.

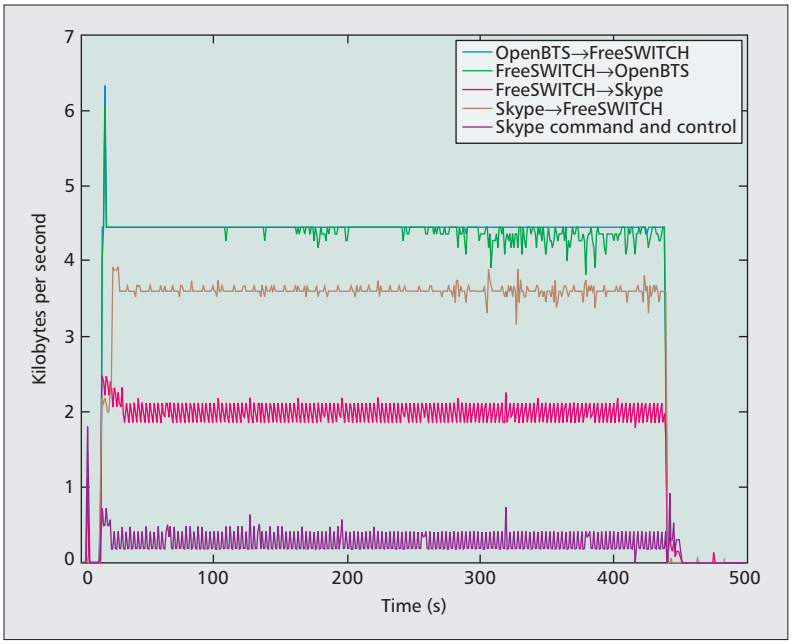


Figure 3. Network utilization for one Skype call.

Rapid Deployment

EmergeNet is designed to be rapidly deployed anywhere in the world. EmergeNet can be deployed as a standalone system including the cellular base station, a solar power system, and a satellite Internet backhaul. Alternatively, EmergeNet nodes can be deployed in coordination with aid organizations, relying on shared or pre-existing power and network infrastructure.

Each EmergeNet base station includes an AC adapter and a DC buck-boost converter to ensure that EmergeNet base stations can be powered from any 100–240 V AC or 4–32 V DC power source capable of providing 30 W of power. This includes vehicle batteries, portable generators, and power grids. To ensure these power connections can be made, each EmergeNet base station includes a 12 V vehicle power port connector, plug adapters compatible with AC outlets in over 150 countries, and battery terminal clamps.

If operation from solar power is desired, each EmergeNet node can include four 100 W solar panels and three 100 Amp-hour 12 V non-spillable sealed lead acid (SLA) batteries. This solar power configuration is designed to provide 24/7/365 uptime in most locations. Due to the low power consumption of EmergeNet nodes (< 30 W), the battery bank will provide a minimum of four days of runtime without sun. The solar array is sized to fully recharge the battery bank with one day of sun in latitudes within 30° of the Equator. The solar power system is modular and can continue operating even if several of the solar panels or batteries are stolen, damaged, inefficiently deployed, or relocated to support other infrastructure. Additional solar panels can be added for deployments in regions further from the Equator, or as weather conditions require.

We selected our batteries with rapid deployability as a key consideration. While many batteries are subject to limitations on transport, non-spillable SLA batteries are not considered hazardous if they comply with the International Air Transport Association Dangerous Goods Regulations Section 4.4, Special Provision A67. This permits EmergeNet nodes to be shipped worldwide with batteries installed. We have traveled with SLA batteries on U.S.-based and European air carriers without incident.

By using Skype to route inbound and outbound calls and messages, EmergeNet supports voice calling and text messaging to and from phones in 170 countries, with no country-specific configuration required. This is particularly important for enabling rapid deployment worldwide. Using Skype to access the public switched telephone network eliminates the need to configure SMS and VoIP interconnects in each country EmergeNet is deployed, and for each country EmergeNet users want to call. Finding reliable and functional interconnects was a primary challenge faced by Heimerl *et al.* in Papua. In this work, researchers evaluated tens of SMS-routing companies, and determined that "an exhaustive search for the correct partner will be required whenever deploying in a new country" [2]. EmergeNet's design eliminates the time- and labor-intensive task of setting up and evaluating VoIP and SMS interconnects.

EmergeNet is designed to be deployed easily by one or two people with limited technical knowledge. With an EmergeNet node in the bed of a pickup truck, a successful deployment should take less than 10 minutes and consist of only a few steps:

- Park the truck facing north, to provide maximum solar radiation for the south-facing solar array in the bed.
- Raise the antenna mast.
- Align satellite antennas or long-range Wi-Fi antennas based on the current location.
- Power on the base station.

EmergeNet nodes can operate independently or in coordination with each other. Each EmergeNet node will periodically check for the existence of other EmergeNet nodes on the local network. If an EmergeNet network is found, new nodes will automatically join the pre-existing EmergeNet network. Otherwise, the node creates its own EmergeNet network. This process requires no human intervention and is further explained in the next section.

Automatic Reconfiguration

A core design goal of EmergeNet is to offer graceful service degradation in place of failure. Each cellular base station will react to problems with other EmergeNet nodes or the local network infrastructure. In systems with more than one base transceiver station (BTS), each BTS evaluates the availability of its neighbors and will reconfigure itself accordingly.

We use monit, a process monitoring daemon, to increase the reliability and robustness of each cellular base station. Monit monitors the availability of the base station's services, as well as the availability of the main BTS and neighboring BTS units. Monit automatically restarts failed software components and proactively prevents the base station from entering error conditions, such as filling the disk or running out of memory. We configure monit to check the status of all services

every 10 s, which is frequent enough to detect failures rapidly, and uses at most 0.2 percent of the base station's CPU and memory. More frequent monitoring may increase writes to disk and network traffic, while less frequent monitoring lengthens average downtime due to failure.

Many community cellular networks, including VillageCell, rely on a central private branch exchange (PBX), such as FreeSWITCH, to route calls and text messages. When this PBX is not available, SIP messages from OpenBTS to FreeSWITCH will not receive responses. Without responses from FreeSWITCH, OpenBTS loses all functionality. Users cannot place calls or send text messages, even to users on the same BTS.

To allow any base station to quickly take over the role of main BTS, each EmergeNet base station is capable of acting as the "main BTS" by launching sipauthserve, smqueue, and FreeSWITCH. We configure OpenBTS to send SIP traffic to a floating IP address, which is not owned by any specific machine on the network. This allows one or more base stations to claim the floating IP and take over the role of main BTS, enabling rapid recovery from base station or network failures. Each base station will periodically check whether the main BTS is available. When the main BTS is unreachable, the base station will check whether higher-ranked base stations are accessible. If they are, the base station waits for one of the higher-ranked base stations to become the main BTS. If no higher ranked base station is available, the base station claims the floating IP address, advertises the route change to the network, and becomes the main BTS. Traffic will automatically be rerouted to that base station. This ensures that even when network connections between base stations fail, each base station continues to route calls and messages locally and to any reachable BTS.

Evaluation

In this section, we explore the suitability of EmergeNet for EDNs, and demonstrate EmergeNet's rapid recovery from failure.

EmergeNet Network Traffic Evaluation

To evaluate EmergeNet's performance on the network level, we performed several calling and messaging experiments. We developed an Android application that programmatically sent messages and placed calls from 10 mobile devices camped on two EmergeNet base stations. We captured all network traffic at each base station using tcpdump. As shown in Fig. 3, each traffic stream was classified based on its endpoints, revealing the traffic impact of each EmergeNet component. We measured the bandwidth consumption of each EmergeNet feature, averaged over 10 calls, registrations, or messages. For messaging tests, we performed each experiment twice, using the minimum (1 character) and maximum (160 characters) SMS message length.

As Table 1 shows, EmergeNet nodes are well suited to operating in EDNs with limited bandwidth backhaul connections and congested local networks. When calling remote users, each Skype call uses only 2 kB/s on the backhaul link for outbound voice traffic and 4 kB/s for inbound voice traffic. Skype background traffic, such as presence notifications and peer selection, consumes less than 1 kB/s per user on average. A standalone EmergeNet system generates no additional local network traffic, since OpenBTS, FreeSWITCH, smqueue, and sipauthserve all operate on the same machine and communicate via the loopback interface. In networks with multiple EmergeNet base stations, communication between EmergeNet components generates small amounts of traffic on the local network, which connects EmergeNet base stations to each other. Voice packets passed between base stations use RTP over UDP. Therefore, packet loss on the local network will impact the audio quality of calls, but EmergeNet nodes can continue operating on lossy networks without negatively impacting the local network with useless retransmissions of audio packets.

	Local network	Uplink	Downlink
Skype call	8.6 kB/s	2 kB/s	4 kB/s
Local call	8.6 kB/s	0	0
SMS to Skype	1.4–1.6 Kb	2.9–3.7 Kb	1.4–1.7 Kb
Skype to SMS	1.4–1.6 Kb	1.1–1.9 Kb	1.7–2.9 Kb
Local SMS	2.8–3.2 Kb	0	0
Registration	1.52 Kb	0	0

Table 1. Bandwidth utilized by EmergeNet activity.

Rapid Recovery

EmergeNet nodes monitor their services and performance in order to maintain the availability of service and decrease downtime. We evaluate EmergeNet's robustness and rapid recovery by inducing an error condition, then measuring the length of time before the error is detected and service restored. For the OpenBTS, FreeSWITCH, and Main BTS failure tests, we consider service restored when an EmergeNet GSM handset can send a Skype chat message to a remote Skype user. We consider smqueue and sipauthserve functional when they respond to SIP messages on their respective ports.

The system takes less than two minutes to go from "off" to "operational." As Table 2 demonstrates, EmergeNet nodes are unlikely to face downtime lasting more than two minutes. The average time to detect a failure in sipauthserve or smqueue is generally just over half monit's 10 second (s) polling interval. Failures in FreeSWITCH could take two polling intervals to be discovered, as the FreeSWITCH process may appear functional while shutting down. Unlike the relatively simple sipauthserve and smqueue, OpenBTS and FreeSWITCH require initialization before they begin functioning properly. This is reflected in the longer recovery time for those components, 20.2 and 24.6 s, respectively. In the event of main BTS failure, the network waits about one minute to permit the main BTS to recover from its failure and restore service. In the worst case scenario where the main BTS cannot recover, a

	Time to detect	Total downtime
OpenBTS failure	< 1 s	20.2 s
FreeSWITCH failure	10.4 s	24.6 s
smqueue failure	6.1 s	6.5 s
sipauthserve failure	6.3 s	6.4 s
Main BTS failure	61.6 s	106.3 s

Table 2. Mean time to detect and correct failures over 10 trials.

second base station takes over the IP address of the main BTS, launches all required services, and becomes the new main BTS. This process takes roughly 45 s from the time failure is detected, resulting in downtime of 106.3 s on average.

Conclusion

EmergeNet provides disaster victims and first responders with a robust, portable, rapidly deployable cellular telephony system. The EmergeNet cellular network is affordable, low-power, and provides voice and text messaging services to wide areas. We utilize the Skype VoIP network to enable free and low-cost calling and messaging to Skype users and phones in over 170 countries, and show that this Skype traffic does not over-utilize EDN backhaul connections or local networks. EmergeNet is an improvement over traditional hierarchical cellular networks in emergency scenarios as it has no reliance on carrier infrastructure, which is often damaged in disasters. We believe EmergeNet's model of self-contained cellular base stations using standard IP networking to connect to each other and the Internet will prove to be extremely robust in emergency and disaster scenarios.

Acknowledgment

This work was supported in part by NSF Network Science and Engineering award CNS-1064821 and an award from the U.S. State Department.

References

[1] ITU, "Measuring the Information Society 2013," http://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2013/MIS2013_without_Annex_4.pdf, 2013.
[2] K. Heimerl *et al.*, "Local, Sustainable, Small-Scale Cellular Networks," *Proc. 6th Int'l. Conf. Info. and Commun. Technologies and Development*, Cape Town, South Africa, Dec. 2013.
[3] Rhizomatica, "First Site Up and Running," http://rhizomatica.org/2013/03/12/first-site-up-and-running, Mar. 2013.
[4] J. Butler *et al.*, *Wireless Networking in the Developing World*, 3rd Ed. Limehouse Book Sprint Team, 2013.
[5] Info. Technology Disaster Resource Center, "Hurricane Sandy Response after Action Report & Recommendations," http://itdrc.org/pubs/whitepapers/Hurricane_Sandy-ITDRC_AAR-Jan2013.pdf, Jan. 2013.
[6] M. Zheleva *et al.*, "Kwiizya: Local Cellular Network Services in Remote Areas," *Proc. 11th Annual Int'l. Conf. Mobile Systems, Applications, and Services*, Taipei, Taiwan, June 2013.
[7] J. Genachowski, "Superstorm Sandy Field Hearing New York, NY and Hoboken, NJ," http://transition.fcc.gov/Daily_Releases/Daily_Business/2013/db0205/DOC-318754A1.pdf, Feb. 2013.
[8] A. Anand *et al.*, "VillageCell: Cost Effective Cellular Connectivity in Rural areas," *Proc. 5th Int'l. Conf. Info. and Commun. Technologies and Development*, Atlanta, GA, Mar. 2012.
[9] M. Zheleva *et al.* "Bringing Visibility to Rural Users in Cote d'Ivoire," *Proc. 6th Int'l. Conf. Info. and Commun. Technologies and Development*, Cape Town, South Africa, Dec. 2013.
[10] D. L. Johnson, E. M. Belding, and G. van Stam, "Network Traffic Locality in A Rural African Village," *Proc. 5th Int'l. Conf. Info. and Commun. Technologies and Development*, Atlanta, GA, Mar. 2012.
[11] *OpenBTS 3.1 Users' Manual*, rev. 11, Range Networks, San Francisco, CA, 2013, pp. 29–31.

Biographies

Elizabeth M. Belding [F] (ebelding@cs.ucsb.edu) is a professor in the Department of Computer Science at the University of California, Santa Barbara (UCSB). Her research focuses on mobile networking, and information and communication technology for development (ICTD). She is the author of over 100 technical papers and has served on over 60 program committees for networking conferences. She is currently on the editorial board of *IEEE Pervasive Magazine* and an Editor-at-Large for *IEEE Transactions on Networking*. She is an ACM Distinguished Scientist.

Daniel Iland (iland@cs.ucsb.edu) is a Ph.D. student in the Department of Computer Science at UCSB. He was awarded a Bachelor's degree in computer science from the Rochester Institute of Technology in 2011. His research focuses on enabling robust and reliable communication in emergency and disaster scenarios, wireless localization, and ICTD.second

Exploiting the Use of Unmanned Aerial Vehicles to Provide Resilience in Wireless Sensor Networks

Jó Ueyama, Heitor Freitas, Bruno S. Faiçal, Geraldo P. R. Filho, Pedro Fini, Gustavo Pessin, Pedro H. Gomes, and Leandro A. Villas

Jó Ueyama, Heitor Freitas, Bruno S. Faiçal, Geraldo P. R. Filho, and Pedro Fini are with the University of São Paulo.

Gustavo Pessin is with the Vale Institute of Technology (ITV).

Pedro H. Gomes is with the University of Southern California.

Leandro A. Villas is with the University of Campinas.

The authors would like to acknowledge the financial support granted by FAPESP (São Paulo State Research Foundation, processes 2014/06330-5 and 2012/22550-0), CNPq (Brazilian Research Council, processes 473493/2013-6 and 132007/2013-4) and Capes (National Council for the Improvement of Higher Education, process DS-6706658/D). Jó Ueyama would also like to thank the Office of Naval Research Global for funding part of his research project.

ABSTRACT

A wireless sensor network is liable to suffer faults for several reasons, which include faulty nodes or even the fact that nodes have been destroyed by a natural disaster, such as a flood. These faults can give rise to serious problems if WSNs do not have a reconfiguration mechanism at execution. It should be noted that many WSNs designed to detect natural disasters are deployed in inhospitable places and depend on multihop communication to allow the data to reach a sink node. As a result, a fault in a single node can leave a part of the system inoperable until the node recovers from this failure. In light of this, this article outlines a solution that entails employing unmanned aerial vehicles to reduce the problems arising from faults in a sensor network when monitoring natural disasters like floods and landslides. In the solution put forward, UAVs can be transported to the site of the disaster to mitigate problems caused by faults (e.g., by serving as routers or even acting as a data mule). Experiments conducted with real UAVs and with our WSN-based prototype for flood detection (already deployed in São Carlos, State of São Paulo, Brazil, have proven that this is a viable approach.

INTRODUCTION

Although in many cases natural disasters cannot be avoided, their effects can be mitigated through issuing warnings and following suitable rescue procedures. In the period following a disaster, the monitoring of affected areas is of vital importance to prevent dangerous measures being taken and to safeguard human lives. With this aim in mind, we have developed and deployed a wireless sensor network (WSN) for urban river monitoring, which is described in details in [1]. Briefly, WSNs are distributed systems composed of sensors that are interconnected through wireless links. They are used to monitor physical phenomena such as temperature, atmospheric pressure, and light exposure for various purposes in, for example, medical, civilian, and military areas. Its form of communication is carried out through multiple hops where nodes communicate with their neighbors until the data reach their final destination.

Our WSN-based system monitors river floods, and issues warnings to the population and vehicles at risk. To date, we have constructed and deployed eight sensor nodes in the city of São Carlos in São Paulo, Brazil. The urban rivers in the city of São Carlos have an elongated shape, and hence the sensors were deployed linearly along the river (Fig. 1). Urban rivers in Brazil are usually surrounded by big roads such as the so-called Marginal Tietê in São Paulo. In São Carlos, there is the Marginal Tijuco Preto road. As a result, these roads tend to get flooded, which means there is a need to send warnings to drivers to avoid those areas when they are at risk.

Since the sensor nodes are deployed linearly, we have adopted ZigBee multihop communication, which enables each node to send its packets to its neighbor toward the sink node. Clearly, the failure of a node will compromise the WSN either completely or partially. In particular, a couple of sensor nodes may be washed away during a flash flood, causing damage to the whole system. We have devised a number of fault tolerance mechanisms to ensure robustness and resilience, including the use of a third generation (3G) network in case the multihop transmission fails. Obviously, not all the sensor nodes have a 3G network as a way of providing a lighter prototype. In this scenario, a couple of sensor nodes with a 3G network may be destroyed, and/or the 3G network might not be operating during the period of a critical flood.

As a result, we propose the use of unmanned aerial vehicles (UAVs) to make the WSN more resilient to failures and natural disasters that our river monitoring WSN is prone to suffer. UAVs are used for various tasks, whether they are civil-

 0163-6804/14/$25.00

The e-NOÉ employs a pressure sensor to undertake this work because one of the basic principles of physics is that the degree of pressure exerted at a particular point within the river depends on the height of the water column above it.

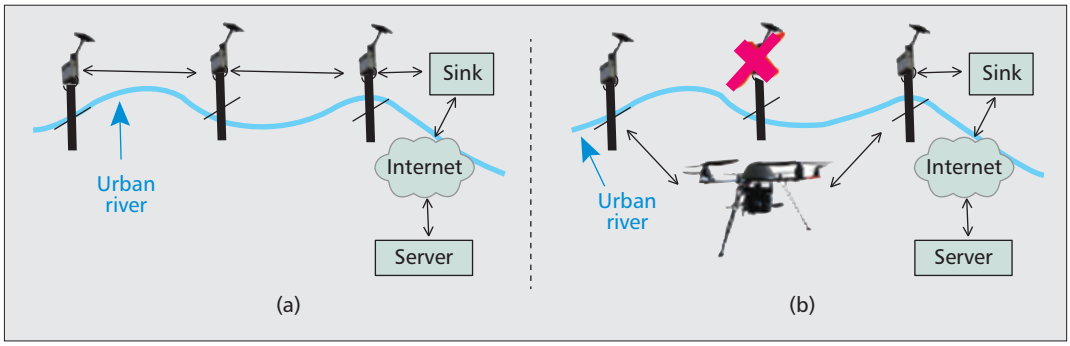


Figure 1. When the WSN is in normal operation mode, the messages with information about the monitoring of the urban river are transmitted over multihop to the sink node: a) the sink node sends all the information to a central processing unit through an Internet connection; b) when a failure occurs in a sensor node, previous nodes cannot transmit their data to the sink node. Since this part is without communication, it remains temporarily inoperable until the failed node has been repaired or connection re-established. By employing our prototype of the mobile node, it is possible to re-establish the connection or collect data throughout the WSN and convey it to the processing center. This choice is carried out in accordance with the plan of action used.

ian or military, including surveillance, reconnaissance, monitoring, and aerial mapping. Another area where UAVs can be employed is to maintain connectivity with a WSN if there are failings in the infrastructure. In our proposal, the UAV Microcopter (an unmanned mini-helicopter with eight propellers) can be employed in two key capacities: i) to act as a router in case a node fails to transmit packets in multihop communication (Fig. 1); and ii) to serve as a data mule [2] for data dissemination, which can help us to form a delay-tolerant network (DTN) in which packets from our WSN can be disseminated to a vehicular ad hoc network (VANET) so that drivers can avoid flooded roads

To show the feasibility of our proposal, an investigative study of wireless communication has been undertaken between terrestrial sensors and the UAVs. The key feature of the analysis is an assessment of the electric power consumption of the wireless communication device, which is fitted to the UAV at the network response time and in the packet losses.

The remainder of this article is structured as follows. We outline related work in the field. We explain the approach that has been adopted. We give a detailed account of the experimental environment. We include a discussion of the results of the experiments, and we summarize the conclusions of the authors about the results of this study.

Related Work

Although there are works such as [3–5] that propose the use of UAVs to integrate WSNs, it was noted that there are no studies that investigate the use of UAVs as a gateway and data mule, or include an analytical study of energy consumption. Moreover, only a few of them use the ZigBee protocol, which is the most widely used standard for WSNs and the focus of our study.

The aim of the study by Hauert and colleagues [3] was to set up a communication network in disaster areas through UAVs. The project employs UAVs to spread out the nodes of a network in a disaster zone with the aim of establishing communication with rescue teams. Some real UAV prototypes were constructed with modules from a global positioning system (GPS) module and a ZigBee transmitter.

The main proposal of Freitas and colleagues [4] is to recommend the use of UAVs to give support/connectivity to the WSNs installed on land. The principal feature is to investigate an attempt to provide connectivity between the UAVs and the subnetworks formed by the WSN nodes on the ground. In carrying out this study, the authors assume that the WSN network was abstracted, and the study is focused on the way the UAVs operate.

The study by Tuna and colleagues [5] outlines strategies for the use of UAVs in implementing a WSN for monitoring a post-disaster recovery phase. These strategies involve determining routes, making improvements to positioning, and finding better sites for placing the wireless sensors. The proposal uses UAVs to spread out wireless sensors in the area coverage of the WSN network. The proposed system was simulated with USARSim to assess the location and navigation performance of a UAV responsible for the layout of the WSN. The study does not specify what kind of wireless technology was used.

As mentioned earlier, the use of a UAV to integrate the UAVs and WSNs is indeed not novel, and there are a few works that address this issue. Our group has just published a paper[1] that outlines the use of UAVs and WSNs for detecting pesticide drifts while spraying chemicals on crop fields. Having said that, it should be stressed that none of the works make use of UAVs to provide a higher degree of resilience for WSNs to act against natural disasters (e.g., floods) or make evaluations based on real devices. For this reason, our work includes real experiments that involve both prototyped UAVs and deployed WSNs (i.e., our work does not include validation through a simulation tool). The tested UAV was equipped by our group with a Zigbee transmitter and a Raspberry Pi so that our communication evaluations could be conducted with the deployed WSN.

[1] B. S. Faiçal *et al.*, "The Use of Unmanned Aerial Vehicles and Wireless Sensor Networks for Spraying Pesticides," *J. Systems Architecture*, vol. 60, no. 4, 2014, pp. 393–404.

The Proposal

This article seeks to show how mobile nodes are capable of providing resilience to a WSN employed for urban river monitoring. The WSN obtains information about the behavior of the river and transmits it to a central processing unit where it can be used for predicting a possible flood. In addition, we have constructed a real prototype and conducted an exploratory analysis of the energy consumption of a UAV and sensor nodes (without the help of a simulation tool). A sensor node that is able to communicate with the WSN deployed in the area of interest (on the bank of an urban river) is coupled to the UAV, and thus makes a mobile node. This node is used when a fault in the WSN makes part of the network inoperable. Although an alternative strategy is to add nodes to the network to increase fault tolerance/resilience, this solution shortens the battery life because it increases the amount of data being transmitted, as well as the cost incurred by the duplication of equipment and deployment. Additionally, this solution does not guarantee fault tolerance/resilience due to the prevailing environment conditions at the time of the disaster. On the other hand, the use of a UAV to provide resilience in the WSN adds some important extra features, such as:

- The UAV can be equipped with a camera to transmit images in real time for rescue teams.
- It can act as a data mule between the points separated by the communicating nodes.
- It can map out the affected disaster areas while performing other activities.
- Unlike vehicles (of the rescue teams), it can navigate over rivers, creeks, and areas affected by floods, landslides, and earthquakes.

In light of this, our UAV will perform two key roles:

- To serve as a router in multihop transmission. While it is unable to last for a long time, we still argue that the critical information will arrive in a timely manner for the population and during the lifetime of the UAV battery. UAVs can also send images (taken by themselves) to rescue teams so that they can mitigate natural disasters such as floods and landslides.
- To act as a data mule. This will help us to construct a DTN-type network. The data about the river, such as its depth level, can be conveyed and subsequently downloaded to the network infrastructure such as that of VANETs. This can be integrated to several existing data dissemination algorithms for VANETs such as the [6].

Figure 1 shows a WSN that is deployed at the edge of an urban river to monitor the behavior of the water flow. The prototype was deployed at this location due to the frequency of the floods there, particularly during periods of torrential or prolonged rainfall. It is thus possible to observe the standard operation of the WSN (Fig. 1a). The messages with information about the urban river are transmitted through multihop to the sink node so that the information about the entire WSN can be sent to the base station. However, WSNs are subject to failures, and this problem renders part of the architecture inoperative, since it cannot transfer the data that has been monitored. Nevertheless, with the use of a UAV acting as a mobile node (Fig. 1b), the new architecture can provide a higher degree of fault tolerance for the entire model.

Predicting Floods: e-NOÉ

A WSN consists of a set of sensor nodes used in an area of interest to monitor different phenomena (e.g., temperature or light exposure) [7]. These sensor nodes can operate as relay networks (repeaters in a wireless network employed to transmit data to a destination outside the transmission range). Each sensor node is basically composed of four units: a sensing unit, a processing unit, communication, and power supply [7, 8]. There are various research studies that employ WSNs [9–11]; prominent among these are studies for the monitoring of urban rivers with the aim of forecasting floods [10, 12].

Our project for monitoring urban rivers is called e-NOÉ, and one of its aims is to detect and predict floods with a view to being able to issue warnings, and avoid serious damage and loss of life. The sensing component in the e-NOÉ project (also called a node) is fitted with sensors (e.g., for pressure and temperature), and these are installed at strategic points along the banks of a river. They are connected to each other through a wireless network. The nodes communicate with each other with the aim of establishing a link with the base station that is responsible for the conversion of the networks.

The e-NOÉ employs a pressure sensor to undertake this work because one of the basic principles of physics is that the degree of pressure exerted at a particular point within the river depends on the height of the water column above it. With a pressure sensor it is easy to measure the height of this column. A sudden increase in the height in a short period of time indicates signs of a possible flood; warnings can then be sent to the people who are in the areas of risk.

The Prototype of a Mobile Node

UAVs are airships that are either capable of carrying out flights autonomously or remote controlled by a base station. This means it is not necessary to have pilots aboard during operations. These airships can undertake various activities in situations of high risk for humans or in areas where it is difficult to obtain access. Compared with conventional airships (flown by pilots aboard), the UAVs offer a safe alternative for various applications at a low cost [13, 14]. The UAVs have a wide range of technological devices that can be of assistance during the flight such as flight controllers, an onboard computer, GPS modules, and sensors. In addition, the UAVs are equipped with wireless communication resources that are able to exchange information with other features of a data network. In this scenario, the UAV becomes a mobile node of this data network.

Two large categories of the various UAV models are worth highlighting and can be classified as follows:

UAVs are airships that are either capable of carrying out flights autonomously or are remote controlled by a base station. This means it is not necessary to have pilots aboard during operations. These airships can undertake various activities in situations of high risk for humans or in areas where it is difficult to obtain access.

Figure 2. Our prototype has the sensor node located in the undercarriage and involves the following items: a) the ZigBee wireless communication module; b) Arduino, used to monitor the power consumption of ZigBee and UAV; c) Raspberry, the computational component of the sensor node; d) the Power Meter Shield, connected to the Arduino and assisting in monitoring the power consumption of ZigBee; e) the temperature and humidity sensors providing basic information on weather conditions; f) the antenna coupled to ZigBee.

- Fixed-wing UAVs — These have a design like traditional planes where the vehicle support is obtained through wings that are fixed in line with the fuselage.
- Those with rotary wings — These are helicopters where the vehicle support is obtained through the wings (propellers), which move in line with the fuselage.

The rotary wing UAVs possess features that are closely linked to the objectives of this study. The ability to take off and land with vertical guidance allows the prototype to initiate and finalize its flights in environments where the ground is irregular. In addition, these models can hover and need less aerial space to carry out maneuvers than fixed-wing UAVs. These features provide a prototype with greater flexibility. Figure 2 shows details of our prototype.

In our prototype, we developed a module that is responsible for the entire communication between the UAV and the WSN, which is located in the undercarriage (Fig. 2). Thus, the prototype has become a mobile node that is able to route messages with the aim of overcoming obstacles (e.g., it acts as a bridge to boost the signal range so that it can address the problem of failures in any fixed sensor node), and it can also be employed as a data mule (e.g., by collecting data from the WSN and sending it to the central processing unit).

Performance Evaluation

An experimental environment was created with the aim of obtaining more reliable data[1] for the experiments. Three agents were used in this scenario: i) the client, ii) the UAV, and iii) the server. The client is responsible for making the requests to the server, as well as coordinating the UAV activities. The UAV carries out the role of routing the packets between the client and the server as well as monitoring the energy consumption of the ZigBee radio module. The server is responsible for replying to requests made by the client.

Three response variables were used for the performance evaluation:

- Energy consumption, for determining how much is being consumed by the ZigBee communication module
- Round-trip time (RTT) delay, which shows the transmission time from the transmitter node to the receiver node when passing through the UAV
- The packet loss rate, which represents the percentage of packets lost during the transmissions

The set of primary factors summarized in Table 1 was selected with the aim of conducting a performance evaluation of the UAVs. In carrying this out, the proposal is evaluated by means of a UAV (both on the ground and in flight), different distances (30 and 80 m) and different antenna gains (3 and 10 dBi). A complete factorial design was employed to enable every combination of the sets of factors to be put into effect. Each set of experiments was replicated 33 times, and a statistical comparison between the sets were carried out by employing the Shapiro-Wilk normality test and the Wilcoxon rank sum test.

Other relevant information gathered during the flight that influenced our performance evaluation is as follows:

- ZigBee channel: 16
- Room temperature: (26.98 ± 0.07) °C
- Relative humidity: (52.15 ± 0.11) percent
- Readings for the rate of energy consumption: 170 readings/s
- ZigBee communication rate: 9600 b/s
- ZigBee transmission power: 10 dBm

Energy Consumption of the Sensor Nodes

The energy consumption with regard to antenna gain and distance can be seen in Fig. 3a. It should be noted that the lowest confidence interval obtained occurred when the UAV was in flight (experiments E5 to E8), which shows a concentration of energy consumption close to the average ~ 2.38 mJ. Larger dispersion was obtained when the UAV was on the ground (experiments E1 to E4). Hence, the wireless communication performance is better when the node is further away from the ground. Information about this is obtained through the propagation of radio waves in the atmosphere, and it is influenced by reflections from the terrestrial soil [15]. It can be concluded that there is a reduction of energy consumption in the ZigBee communication module when the UAV is in flight.

Energy Consumption of the UAV

Figure 3b shows a sample with different conditions for the energy consumption of the UAV in flight. These conditions, when divided into regions, are described and discussed as follows:

- *Flying over with low wind*. The observed condition in this area is caused by an air passage. We detected a period of instability from the time when a current of air came into contact with the Microcopter. This is offset by an increase in torque and hence

[1] Data of the experiments are available at http://goo.gl/glUjUh.

the rotation of engines. This situation results in an increase in the energy consumption of the system.

- *Flying over with no wind — an ideal condition for the Microcopter*. In the ideal condition, the Microcopter tends to keep a state of equilibrium and only has to maintain a fixed altitude. In this case, a reduction in the power exerted by the engines can be observed; hence, there is a decrease of energy consumption.
- *An increase in the rotation of the motors to reach a higher altitude*. A considerable dissipation of the power of the motors is required for a sharp rise in altitude. As seen in Figs. 3b and 3c, this maneuver produces peak power.
- *Turning off the engines*. By default, when the motors are turned off, they suffer from an increase in speed, which causes a slight variation in power output.
- *Engines off*. In this situation, the power dissipated by the engines is minimized.

Round-Trip Time Delay

Figure 3c shows the results of the RTT with regard to the type of antenna and distance. It should be noted that the results are statistically equivalent for all cases observed (p-value above 0.05, Wilcoxon rank sum test). The extreme values shown in the RTT results follow a logical pattern since the environmental conditions are unsuitable and there is interference caused by the signals; the reason for this is that WiFi and ZigBee [16] use the same unlicensed frequency band of ~2.4 GHz. Different studies have shown that this feature affects the performance of ZigBee. In the analysis of extreme values, values higher than ~450 ms were not taken into account. It can be concluded that there was no loss or gain to the transmission times in the observed conditions.

Packet Losses

The results of packet losses were 0 percent lost packets for all eight experiments conducted (E1 to E8). It is thought that this was due to the data transfer rates employed. The transfer rate for wireless transmissions was 250 kb/s, and the transmission rate between the ZigBee and the computer was 9.6 kb/s, which was 26 times slower.

Discussion

In accordance with the aims of this article, the prototype showed it could provide mobility for the support node during the experiments in an appropriate way. The sensor node was able to temporarily replace the faulty sensor node in communicating with other sensor nodes without causing delay or packet loss. Moreover, these results suggest the need for a configuration with a more powerful antenna because this can provide a greater communication range and lower power consumption when away from the ground.

On the other hand, the power consumption of the UAV indicates the need for an action plan where the prototype is used as a data mule; in other words, to collect data on the inoperative party and transport them to a central processing unit. It should be noted that the prototype does not have to fly over the exact area where the sensor nodes are fixed, since it may be possible to use an antenna with a longer range. This plan of action has another positive aspect: the prototype often returns to the processing unit, which means its battery can be charged when necessary.

Experiments	Sensor node	Distance	Antenna gain
E1	Ground	30 m	3 dBi
E2	Ground	30 m	10 dBi
E3	Ground	80 m	3 dBi
E4	Ground	80 m	10 dBi
E5	Flight (5 m)	30 m	3 dBi
E6	Flight (5 m)	30 m	10 dBi
E7	Flight (5 m)	80 m	3 dBi
E8	Flight (5 m)	80 m	10 dBi

Table 1. The set of primary evaluated factors.

One drawback of the developed network is its bandwidth, since it operates by relying on ZigBee technology. The network has a data rate that is limited to the ZigBee communication rate, which is at most 115,200 b/s. With this transmission rate, it is not feasible to transmit creek/river images, which is why our network mainly transmits text data (i.e., no images are sent out through the ZigBee network). However, the use of the ZigBee radio is recommended for the context of our application, since it is a radio transmitter with long range and low power consumption.

Finally, it is worth remembering that this methodology seeks to explore the energy consumption of sensor nodes of WSNs used to predict floods and validate the proposed prototype of the mobile node. In this way, it can lead to the development of an action plan aimed at providing resilience for this architecture.

Conclusion

This article outlines the use of UAV as a means of providing robustness and resilience for wireless sensor networks. We conduct a review of related work and discuss the literature in the field. We also outline our proposal and the two key roles that the UAV can play in the event of node failures with our existing WSN. Experiments were carried out with a real UAV and an existing sensor node for river monitoring. Our WSN prototype is able to detect and predict; but when it is deployed in critical areas (i.e. where floods take place), it becomes highly prone to natural disasters, such as a flood itself.

In carrying out its task as a router in a WSN, the UAV has proven effective in terms of energy consumption, which makes it suitable for operating in controlled flight conditions and thus

ensures the effectiveness of network operations in applications for flood detection. We conducted experiments with RTT to measure vertical transmission time from the UAV to sensor nodes. This is of particular interest when using UAV as a data mule. In addition, we ensured that the RTT times did not undergo variations in vertical transmission flow and the eight propellers of the UAV. We also made sure that there were no packet losses, so that the reliability features of ZigBee for the use of applications such as the e-NOÉ WSN-based river monitoring system could be validated.

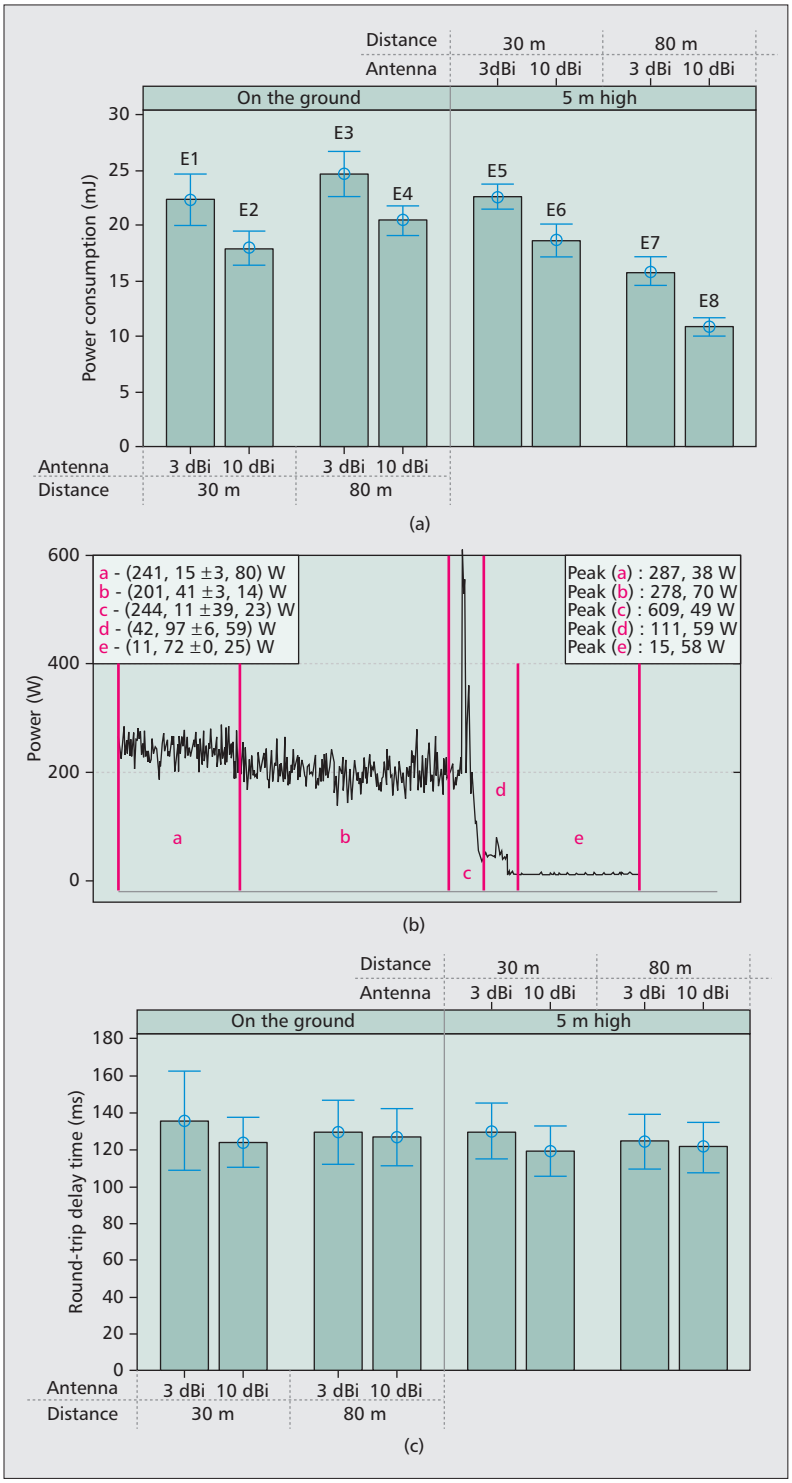


Figure 3. Results: a) energy consumption; b) power consumption of UAV during real flight; c) RTT delay.

To recap, our work included real experiments with both a prototyped UAV and a deployed WSN for flood monitoring (i.e., our work does not include validation through a simulation tool). In addition, it is important to point out that the tested wireless communication technology was able to transmit packets under an adverse condition (i.e., torrential rain). Due to that, the deployed WSN could predict three floods in the city of São Carlos. Our UAV was equipped with the same wireless communication technology; therefore, we believe that our UAV can help us to restore broken wireless communication and also serve as a data mule for providing a WSN with a higher degree of resilience against disasters.

The next stages of this project will be as follows:

- Other antennas with different ranges will be evaluated to minimize the route of the prototype.
- Evolutionary methods will be employed for route planning of action missions for a prototype that can be used as a data mule.
- The prototype will be evaluated in different weather conditions.

References

[1] D. Hughes *et al.*, "A Middleware Platform to Support River Monitoring Using Wireless Sensor Networks," *J. Brazilian Computer Society*, Springer, vol. 17, no. 2, 2011, pp. 85–102.

[2] O. Tekdas *et al.*, "Using Mobile Robots to Harvest Data from Sensor Fields," *IEEE Wireless Commun.*, vol. 16, no. 1, Feb. 2009, pp. 22–28.

[3] S. Hauert *et al.*, "The Swarming Micro Air Vehicle Network (Smavnet) Project," http://lis2.epfl.ch/CompletedResearchProjects/SwarmingMAVs/, 2010, accessed Nov. 30, 2012.

[4] E. P. de Freitas *et al.*, "UAV Relay Network to Support WSN Connectivity," *2010 Int'l. Congress on IEEE Ultra Modern Telecommunications and Control Systems and Wksps.*, 2010, pp. 309–14.

[5] G. Tuna *et al.*, "Unmanned Aerial Vehicle-Aided Wireless Sensor Network Deployment System for Post-Disaster Monitoring," *Emerging Intelligent Computing Technology and Applications*, Springer, 2012, pp. 298–305.

[6] L. Villas *et al.*, "Network Partition-Aware Geographical Data Dissemination," *2013 IEEE ICC*, June 2013, pp. 1439–43.

[7] I. F. Akyildiz and M. C. Vuran, *Wireless Sensor Networks*, Wiley, 2010.

[8] I. Akyildiz *et al.*, "A Survey on Sensor Networks," *IEEE Commun. Mag.*, vol. 40, no. 8, Aug. 2002, pp. 102–14.

[9] G. P. R. Filho *et al.*, "Nodepm: A Remote Monitoring Alert System for Energy Consumption Using Probabilistic Techniques," *Sensors*, vol. 14, no. 1, 2014, pp. 848–67.

[10] J. Ueyama *et al.*, "Applying a Multi-Paradigm Approach to Implementing Wireless Sensor Network based River Monitoring," *2010 First ACIS Int'l. Symp., Cryptography and Network Security, Data Mining and Knowledge Discovery, E-Commerce: Its Applications and Embedded Systems*, 2010, pp. 187–91.

[11] L. A. Villas, *Data Aggregation, Spatio-Temporal Correlation and Energy-Aware Solutions to Perform Data Collection in Wireless Sensor Networks*, Ph.D. dissertation, Federal Univ. Minas Gerais, Mar. 2012.

[12] D. Hughes *et al.*, "GridStix: Supporting Flood Prediction Using Embedded Hardware and Next Generation Grid Middleware," Int'l. Symp. on a World of Wireless, Mobile and Multimedia Networks, 2006.

[13] H. Xiang and L. Tian, "Development of a Low-Cost Agricultural Remote Sensing System Based on an Autonomous Unmanned Aerial Vehicle (UAV)," *Biosystems Eng.*, vol. 108, no. 2, 2011, http://dx.doi.org/10.1016/j.biosystemseng.2010.11.010, pp. 174–90.

[14] L. Villas *et al.*, "3D Localization in Wireless Sensor Net-

works Using Unmanned Aerial Vehicle," *12th IEEE Int'l. Symp. Network Computing and Applications*, 2013.
[15] N. Ya'acob *et al.*, "Observation of Tweek Characteristics in the Midlatitude D-Region ionosphere," *2012 IEEE Symp. Wireless Technology and Applications*, Sept. 2012, p. 28.
[16] Digi, "Xbee/Xbee-Pro Module Product Datasheet," http://ftp1.digi.com/support/documentation/90000976_G.pdf, 2012, accessed on May 10, 2012.

Biographies

Jó Ueyama (joueyama@icmc.usp.br) completed his Ph.D. in computer science at the University of Lancaster, United Kingdom, in 2006. He is currently an associate professor at the Institute of Mathematics and Computer Science, University of São Paulo (ICMC/USP), and he is also a Brazilian Research Council (CNPq) fellow. His main research interests include computer metworks and distributed systems.

Heitor Freitas (heitorfv@icmc.usp.br) is an M.Sc. student in ICMC/USP. His research interests are unmanned aerial vehicle and wireless sensor networks.

Bruno S. Faiçal (bsfaical@icmc.usp.br) is a Ph.D. student in ICMC/USP. His research interests are wireless sensor networks, unmanned aerial vehicles, machine learning, and bio-inspired computing. He received his M.Sc. from ICMC/USP in 2012

Geraldo P. R. Filho (geraldop@icmc.usp.br) is a Ph.D. student in ICMC/USP. His research interests are wireless sensor networks, smart grids, and machine learning. He received his M.Sc. from ICMC/USP in 2014.

Pedro Fini (pedrofini@usp.br) is a computer engineering undergraduate student at ICMC and the São Carlos School of Engineering (EESC) of USP. At ICMC/USP, he has been working on several research projects related to unmanned aerial vehicle and wireless sensor networks.

Gustavo Pessin (gustavo.pessin@itv.org) is an assistant researcher in the Vale Institute of Technology's Applied Computing Lab. His research interests are mobile robotics, machine learning, and sensor networks. He received his Ph.D. in computer science from ICMC/USP.

Pedro H. Gomes (pdasilva@usc.edu) is a Ph.D. student in the University of Southern California's Autonomous Network Research Group. His research interests are wireless sensor networks, the Internet of Things, network optimization, and multichannel access control. He has an M.Sc. in computer science (2011) from the University of Campinas.

Leandro A. Villas (leandnro@ic.unicamp.br) is an assistant professor in the Institute of Computing of the University of Campinas. He received his Ph.D. in computer science from the Federal University of Minas Garais in 2012. His research interests include context interpretation, distributed algorithms, routing algorithms, wireless sensor networks, and vehicular ad hoc networks.

Network Virtualization for Disaster Resilience of Cloud Services

Işıl Burcu Barla Harter, Dominic A. Schupke, Marco Hoffmann, and Georg Carle

ABSTRACT

Today's businesses and consumer applications are becoming increasingly dependent on cloud solutions, making them vulnerable to service outages that can result in a loss of communication or access to business-critical services and data. Are we really prepared for such failure scenarios? Given that failures can occur on both the network and data center sides, is it possible to have efficient end-to-end recovery? The answer is mostly negative due to the separate operation of these domains. This article offers a solution to this problem based on network virtualization, and discusses the necessary architecture and algorithm details. It also answers the question of whether it is better to provide resilience in the virtual or physical layer from a cost effectiveness and failure coverage perspective.

INTRODUCTION

The way people communicate and do business today is changing. Beyond calling people, we send messages or emails. We upload pictures and videos or post about what we are doing. These services are generally provided by servers located in large data centers. Previously, many companies had various servers located in different locations, but now they outsource their IT services to cloud providers or locate them in private clouds within their company network. As a result, today's communication infrastructures consist not only of communication networks but also storage and compute elements located in big data centers that constitute cloud infrastructures. Even the communication networks themselves will depend on clouds in the near future. Software defined networking (SDN) and network virtualization technologies enable network functions virtualization, where the basic idea is to locate the network elements' intelligence in the cloud and enable the use of standardized proprietary hardware within the networks.

In a nutshell, the networks need the cloud to function, and the clouds need the network for information exchange and especially to reach the end customers. Such interdependence requires conscious coordination between the network and cloud domains. However, these domains are currently often operated by separate entities, making coordinated failure coverage and end-to-end optimization largely impossible. However, to provide sufficient quality of service (QoS) and reliability to customers, services need to be optimized in an end-to-end fashion. Reliability plays a crucial role in the decision to adopt cloud services by businesses and is their primary concern according to a survey conducted of over 3700 companies worldwide [1]. Performance ranks third in the list of concerns and has about the same significance as the second, security. Performance concerns are understandable since service degradation and outages can be mission-critical or even fatal. Outages do happen: in the past two years, there have been many outages, some lasting for hours or days, even occurring in the networks and data centers of governments, cities, airline systems, big cloud, and network providers, affecting many businesses and millions of users [2]. Besides local causes of outages caused by power outages, fiber cuts, server or router failures, and so on, some outages can affect a large area and have an even larger impact on businesses and society (e.g., in natural disasters). Communication network and cloud providers need fast and efficient means for recovering from both localized outages and major disasters. Such mechanisms exist today, but when coupled with the problem of separate operation of cloud and network domains, end-to-end recovery is mostly impossible. This in turn leads to unavoidable outages and/or suboptimal solutions. One way to overcome this problem is network virtualization with combined control of network and cloud resources.

Network virtualization is seen as a key enabler of future Internet and future networks. It decouples services from the underlying physical infrastructure. All the parts of the physical infrastructure (the network links, nodes, and servers) are virtualized. Each network resource or server can host multiple virtual resources simultaneously, which are rented to different service providers, enabling more efficient use of physical resources. An isolated complete virtual network contains these different virtual resource types, where isolation enables the use of a

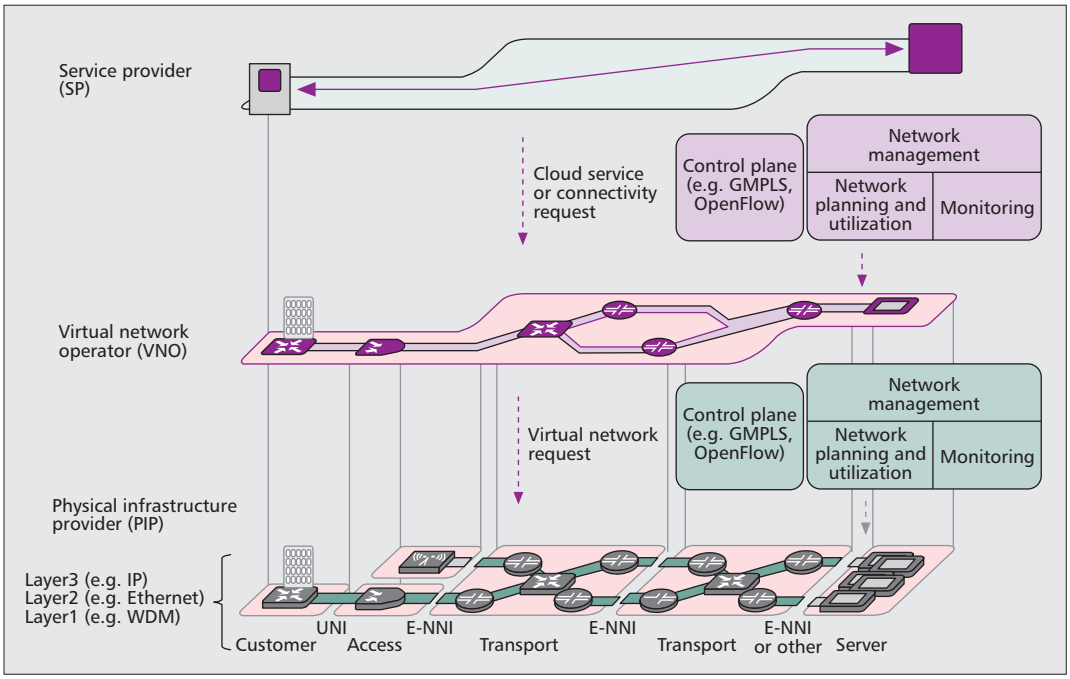
Isıl Burcu Barla Harter is with NOKIA and Technische Universität München.

Dominic A. Schupke is with Airbus Group Innovations.

Marco Hoffmann is with NOKIA.

Georg Carle is with Technische Universität München.

0163-6804/14/$25.00 © 2014 IEEE



A combined control of virtualized network and IT resources is used enabling an end-to-end design and recovery for cloud services, regardless of whether they belong to various PIPs or heterogeneous networks. The last business role is the Service Provider who requests a cloud or connectivity service from the VNO.

Figure 1. Network virtualization architecture showing an example scenario including a service provider (SP), a virtual network operator (VNO) network, and the physical infrastructure of one or more physical infrastructure providers (PIPs) connected via user network interfaces (UNIs) and external network–network interfaces (E-NNIs).

unique layer-specific address space, protocol stack, routing, and QoS definitions. Virtual networks mimic the whole functionality of a physical network, and on top of that offer more flexibility in network design due to an overview of different physical network and cloud domains.

In this article, we propose resilient network virtualization as an approach to disaster recovery, and first describe the network virtualization architecture enabling end-to-end resilience for cloud services. Then we answer the questions of how to design resilient virtual networks and at which layer to apply resilience. We consider different alternatives and compare them in terms of their cost and failure coverage to provide a handy framework to future network providers when deciding on their resilience design. This article extends our previous works [8, 10] by introducing the architectural details and hybrid resilience models, and providing an overview of cost and failure coverage comparison.

Network Virtualization Architecture

In a network virtualization environment, new business roles are expected to emerge [3, 4]. In our architecture, we define three main business roles, as shown in Fig. 1. The physical infrastructure provider (PIP) is the owner of the physical infrastructure, which can consist of fixed or mobile networks (layer 1, 2, or 3) and IT resources like compute and storage, or any combination of them. The physical infrastructure can be composed of multiple PIP domains. The choice of technology in the communication network is not limited; it can be wavelength-division multiplexing (WDM), Ethernet, IP, and so on. A PIP can fully control and monitor its resources, where it can use a generalized multiprotocol label switching (GMPLS) control plane or an SDN-based approach like OpenFlow (OF). A data center PIP is expected to have its own data center network with various interconnected servers. The interface between the data center and WAN depends on the technologies used on both sides. For example, if MPLS is used in the data center, one can easily connect it to the GMPLS WAN with, say, hierarchical label-switched paths (LSPs) or LSP stitching. If OF is used in the data center, the OF controller can communicate with other OF controllers and with GMPLS. For non-MPLS IP virtual private networks (VPNs) and IP overlays not based on VPN like virtual extensible LAN (VXLAN) in the data center, the connection can go over an autonomous system border router (ASBR) and a data center gateway (GW). In the case of an ASBR, there are different options like back-to-back virtual routing and forwarding (VRF), and External Border Gateway Protocol (EBGP) redistribution of labeled VPN-IP routes between neighboring autonomous systems (ASs) without and with multihop EBGP redistribution of labeled VPN-IP routes between source and destination ASs, listed in increasing scalability and decreasing security order [5]. For GW solutions, network overlay stitching can be applied using a data center–WAN GW performing, for example, VRF termination or translation between the virtual network IDs on the data center side and VPN labels on the WAN side [5].

The resources of the PIPs are virtualized and appropriately advertised to the virtual network operators (VNOs). These resources can be virtu-

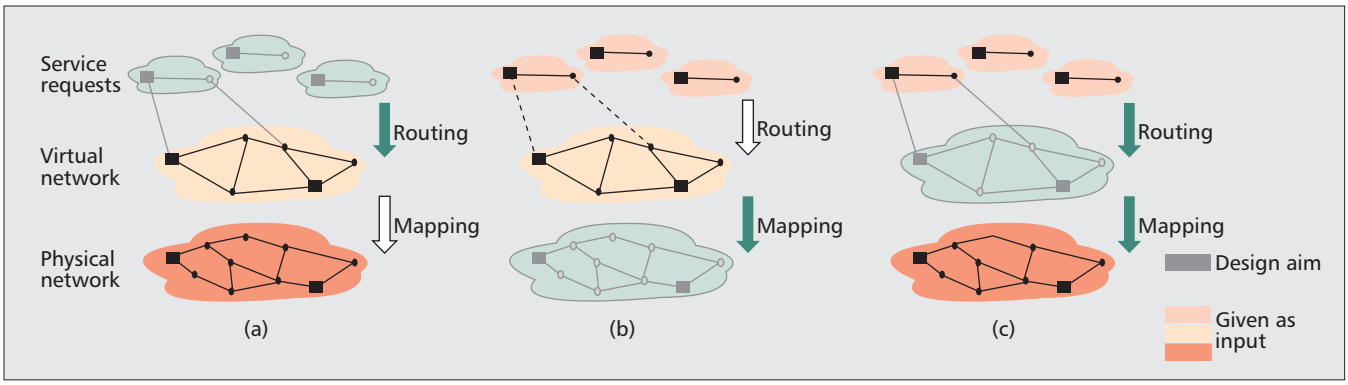


Figure 2. The differentiation between network virtualization, overlay networks, and survivable VPNs: a) in overlay networks, the virtual network and its mapping onto the physical substrate are given as input to the problem and the services need to be routed under survivability constraints; b) for survivable VPNs and virtual network embedding, the virtual network topology is given again and needs to be embedded into the physical infrastructure in a survivable way; c) in network virtualization, however, the virtual network topology is generally unknown a priori. Therefore, it is not taken as input, but has to be determined according to the available physical resources and incoming service requests, which need to be routed in this virtual network.

al network links and nodes as well as virtual machines inside servers. A VNO selects the resources it requires and requests the setup of a virtual network with these resources from the PIP(s). Once the virtual network is established, the VNO has full control over it using its own control and management plane. Combined control of virtualized network and IT resources is used, enabling an end-to-end design and recovery for cloud services, regardless of whether they belong to various PIPs or heterogeneous networks. The last business role is the service provider (SP), who requests cloud or connectivity service from the VNO.

The literature [3, 4] usually defines an additional role, the central broker between many PIPs and VNOs, which we assume to be included in the VNO role since it does not provide an additional effect on the resilience analysis.

Resilient Virtual Network Design: Problem Statement

The aforementioned network virtualization architecture leads to the question of how to design virtual networks for end-to-end-resilient cloud services. Virtual networks are generally similar to overlay networks and VPNs, but there are some differences. In network virtualization, there is a complete isolated network slice as opposed to mere traffic isolation as in VPNs and just node virtualization in the case of overlay networks, which allows the VNOs to operate their service-tailored networks.

Moreover, the design proposals from the VPN or overlay network literature cannot be applied directly. As shown in Fig. 2, in a virtual network environment, the virtual network is mapped on a physical infrastructure, and service requests are routed within the virtual network. Figure 2a shows the case of overlay networks, where the virtual network is already given and the mapping is known. This type of literature addresses how to route the services in a resilient way [6]. For survivable VPNs or virtual network embedding [7], the virtual network is given and should be embedded onto the physical infrastructure in a survivable way, as shown in Fig. 2b. However, a VNO, which needs to design a virtual network to serve its customers, does not have a priori knowledge of a cost-optimal topology. Since a VNO needs to pay a certain fee for renting the virtual resources, it tries to design a virtual network that best fits the requirements of the service requests at a lowest possible cost using input from its customers and the SPs, and knowledge about the advertised resources of different PIPs, as shown in Fig. 2c. A PIP's aim is to serve as many customers, VNOs, as possible, hence efficiently using its physical resources. In order to achieve this, a PIP can favor advertisement of certain virtual resources to a VNO. As a result, we propose the use of a variety of customized virtual network planning and optimization algorithms for a VNO, which can be selected and used according to its needs in order to design resilient virtual networks.

These algorithms can rely on integer linear programs or heuristics. Optimization objectives include minimum cost virtual network design, minimum latency of the service requests [8], and fulfillment of specific QoS requirements while keeping cost in an acceptable range. Special protection mechanisms like shared protection can create win-win situations [9]. It lowers the virtual network setup cost for VNOs by sharing redundant virtual resources among different services. For PIPs, it increases the physical resource usage efficiency and hence enables more customers to be served.

The general structure of these algorithms is described in the following. The algorithm takes as **input**:

- Advertised network resources from the PIP(s) modeled as an undirected physical network graph
- Available data center resources with their network connection nodes
- Set of virtual link and node candidates given as a multigraph connecting all service source nodes with each other and with all possible data center locations

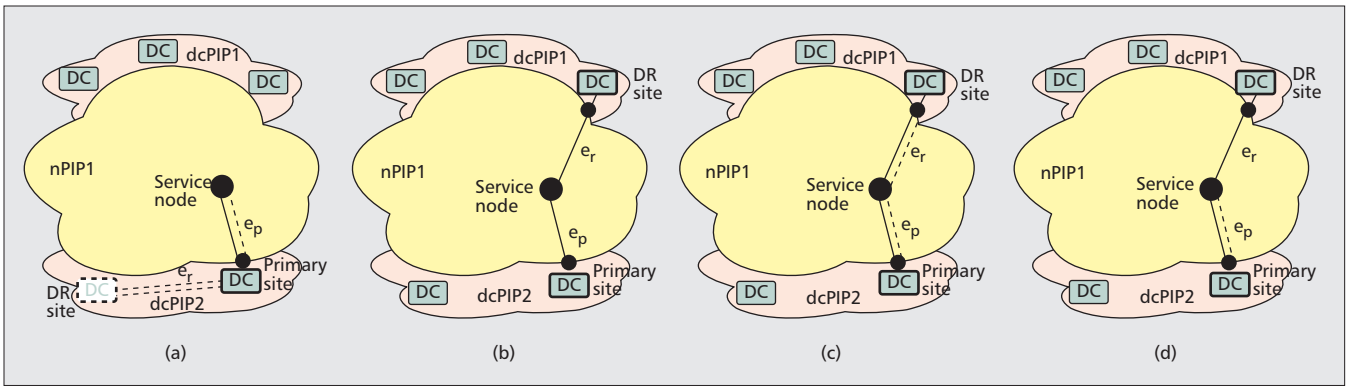


Figure 3. Resilience design alternatives for virtual networks providing protection against network and complete data center (DC) failures: a) Network resilience is provided by using 1:1 protection mapping for the virtual links. The services are routed on a single path in the virtual layer, e_p, to the primary DC site; if it fails, they are routed to the disaster recovery (DR) site in the physical layer on the protection path e_r. This path can be an internal connection of the dcPIP or leased from an nPIP. The DR site and the path e_r are transparent to the VNO; b) Both network and DC resilience are provided by the VNO. The services are routed to two DC locations in the virtual layer, which can belong to different dcPIPs, as opposed to PIP-Resilience. The paths e_r and e_p are physically disjoint; c) DC resilience is provided by the VNO, and network resilience is delegated to the PIP, where in both paths e_r and e_p resilient links are used; d) HPP is similar to HAP with the difference that only the primary path e_p is protected.

- Set of anycast (unicast) service requests, which are defined, for example, by their source node (source and target nodes), network bandwidth, and node resource requirements

The information exchange about the virtual links, nodes, and virtual machines depends on the agreements of the VNOs and PIPs and on the PIPs' business strategy. To enable resilience design, it should typically contain the cost of each element, maximum available capacity, properties of the elements (e.g., end-to-end latency for a virtual link, CPU and memory for a virtual machine) or QoS classes corresponding to certain levels of properties and disjointness information of the virtual elements. A PIP is expected not to disclose its topological information, but to declare if two given virtual links/nodes are physically disjoint or, similarly, if two data centers share any common geographical risks.

The objective is to find a resilient virtual network topology with attached data centers with, say, a minimum virtual network setup cost **such that**:

- The requirements of all service requests are satisfied using physically disjoint routes leading to their primary and disaster recovery (DR) sites.
- The amount of requested resources is within the limit of available virtual and physical resources.

Additional constraints can be used to include specific QoS requirements or different resilience mechanisms like shared protection as mentioned above.

At Which Layer Should Resilience be Provided?

One question when designing resilient virtual networks is at which layer resilience mechanisms should be applied. There are three basic alternatives: providing resilience in the virtual layer by the VNO (VNO-Resilience), in the physical layer by the PIPs (PIP-Resilience), or a combination of both. Resilience in a certain layer has its advantages and drawbacks. The decision metric can vary depending on the priorities of a network provider; these can be, for example, virtual network setup cost, failure coverage, service latency, recovery time, and network utilization. We focus on the first two because there is a trade-off between the price one needs to pay and the offered protection level. Moreover, cost is not the only but usually the main driver for decision making in businesses.

Resilience Design Alternatives

The resilience design alternatives we address in this section are shown in Fig. 3. In each option, one primary and one DR data center site is used for each service. Network resilience, using 1:1 protection, is also provided for the paths leading to those sites. The figure uses a single service as an example to describe the models; however, multiple services are normally routed using multihop routing within the same virtual network, and each service can be routed to any two geographically disjoint data centers. Note that the protection level can be increased by using a higher number of DR sites and a higher level of network resilience, accordingly.

In *PIP-Resilience*, both the network and cloud resilience are delegated to the PIP(s). Each service is routed to a single data center site using a single path within the virtual network as shown with a bold line in Fig. 3a. Since the information about the services is not available at the PIP level, resilience is provided at the virtual link level by using a 1:1 protection mapping for them in the physical layer. For anycast services, cloud resilience is the responsibility of the cloud provider owning the primary data center site. In case of a failure, it redirects the traffic to the DR site in the physical layer. This approach is based on the literature on resilient anycast routing [10]; that is, the services are routed to any two sites, which operate as primary and protec-

tion sites and fulfill the service requirements, with one difference being that the optimization objective is the cost of the virtual network. This recovery action is transparent to the VNO.

If a VNO wants to provision resilience in the virtual layer, it can do so by routing each service to two disjoint data center locations, where the working and protection paths leading to these locations need to be physically disjoint. The same is valid for the unicast case, where the destination nodes of the two paths are identical. This model is called VNO-Resilience and is shown in Fig. 3b. In this case, it is sufficient to have a single path mapping for the virtual links. Moreover, cloud resilience is not limited to a single cloud provider, and the VNO can select any two geographically disjoint data center locations from any provider best suiting the needs of the cloud service requests.

The mathematical formulation of PIP-Resilience and VNO-Resilience models can be found in [8]. The hybrid models are based on the VNO-Resilience model. The main idea behind the usage of the hybrid models is making use of the flexibility of the VNOs in choosing the data center sites and delegating network resilience to the PIPs, which already possess this knowledge and have access to all physical network information. This is a realistic use case for business roles possessing data center resources but no network resources and no network resilience knowledge. Moreover, another big advantage of hybrid models from an operational point of view is the avoidance of unnecessary data center switching due to network failures, which can happen more frequently than complete data center failures.

In the first hybrid model, hybrid all paths protected (HAP), the virtual links used in the paths leading to both the primary and DR sites are resilient, as shown in Fig. 3c. The difference of this model with VNO-Resilience is that there is no longer any need for diversity constraints for the network resources, since network resilience is delegated to the PIP(s).

In HAP, the additional protection against joint data center and backup path failures compared to VNO-Resilience might increase the virtual network price. If failures of the primary data center and the protection path are assumed to be independent, it is sufficient to use unprotected virtual links for the protection path as shown in Fig. 3d, which is called hybrid primary path protected (HPP).

Failure type	VNO		PIP	
	Failure detection	Recovery	Failure detection	Recovery
Transport link failure	Implicit detection	Yes	Yes	Yes
Router/switch/ server failure	Implicit detection	Yes	Yes	Yes
Virtual link failure	Yes	Yes	No	No
Internal virtual machine failure	Yes	Yes	No	No
Complete virtual machine failure	Yes	Yes	Yes	No
Hypervisor (management of VMs) failure	Implicit detection	Yes	Yes	Yes
Control plane (CP) failure	Its own CP	Its own CP	Its own CP	Its own CP
Complete data center failure	Yes	Yes	Yes	Only if it has more than one data center
Sub-network failure	Yes	Yes	Yes	Only if some part of its domain is still intact

Table 1. Possible failures in a virtual network environment, layers they are detectable, and layers that are responsible for the recovery.

Failure Detection and Recovery for Different Business Roles

When deciding on a resilience alternative, the type of potential failures is one of the main considerations. In this section, we briefly list possible hardware and software failures in a virtual network environment and then discuss which of the business roles is in a position to detect them and recover from them.

Table 1 lists the different failure scenarios. We start with the most common failure type in transport networks, physical link failures. A PIP is the owner of the physical infrastructure, and can therefore detect and recover from the physical link failures. Since it is closer to the origin of the failure and since a physical link is usually shared among different virtual networks, a PIP can offer fast and scalable recovery. A VNO is in the position of implicitly detecting a link failure, meaning that it recognizes the failing connection inside its virtual network but cannot detect its actual cause. However, it can apply recovery actions like rerouting the traffic within its virtual network. It has more flexibility due to its overview of different PIP domains while selecting the new route; however, such a recovery action must be taken by every affected VNO separately. The detection of and recovery from a physical node failure is analogous to the case of physical link failures.

In a virtual network environment, another type of link failure is virtual link failure, signifying that the virtual link interface fails. Since the virtual interface failure is an internal failure of the virtual router, a PIP is not in a position to detect it and hence cannot offer recovery from it. The VNO needs to address this problem and can apply a similar recovery action as in the case of a physical link failure. Moreover, a general internal virtual machine failure at the network or server side such as a software problem or buffer overflow can be only detected and solved at the VNO side, except for a complete virtual machine failure that can also be recognized by a

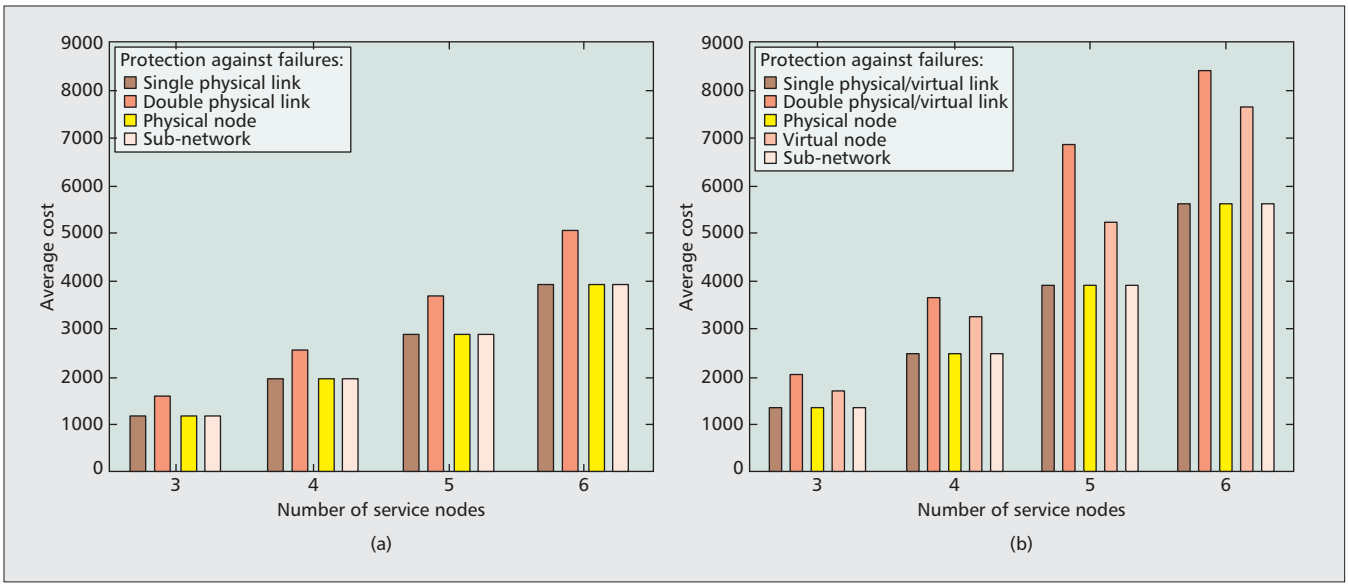


Figure 4. Failure coverage vs. virtual network setup cost for a) PIP-Resilience; b) VNO-Resilience for unicast services showing the principal effects in the comparison of different layer resilience (from our work in [10]).

PIP. Still, normally it is the responsibility of a VNO to restart its virtual machines and take the necessary recovery actions.

In case of a hypervisor failure, which is similar to a physical link/node failure, both roles can detect the failure and recover from it; however, to solve the cause of the problem is the responsibility of the PIP. In the case of a control plane failure, each layer can detect the problems within its own control plane and react to them only. However, since in that case the data plane continues to work and hence a fast recovery is not required, we do not go into more detail on this problem.

Finally, protection against complete data center failures or subnetwork failures, or disaster recovery, can be provided by both business roles. In both of these failure types, where a part of a physical domain or a complete domain is affected, PIPs might have a disadvantage compared to VNOs, who have an overview of different PIP domains. For example, if a PIP only possesses a single data center or the complete PIP domain goes down, it has no chance of offering any recovery for the failed services. However, a VNO can make use of the other available network and cloud domains, and can even have a solid disaster recovery strategy by selecting its resources in advance from disjoint physical domains or availability regions. Availability regions are ideally predetermined such that a failure in one region does not affect the other regions.

In conclusion, recovery against physical failures can be provided by both business roles, where problems occurring within the virtual layer can only be detected and reacted to by the VNOs. Therefore, for physical failure protection, one can choose to provide resilience in the physical or virtual layer, or a combination of both. A recovery strategy in the virtual layer requires reserving redundant virtual resources in advance or requesting them in case of failure depending on the level of protection required, increasing the cost and level of necessary network management knowledge at the VNO. A PIP layer can cope better with physical failures but is restricted in terms of accessing the resources of other domains. Since it is not trivial to decide on the layer to provision resilience, this issue is discussed further in the next section.

What Level of Resilience Should be Used? At Which Layer Should It Be Applied?

Resilience provisioning increases the overall network cost; however, it also increases the service quality and customer satisfaction. Therefore, there is a trade-off between cost and the level of protection or failure coverage an operator should provide.

Cost vs. Failure Coverage

First, we give some insight on how the virtual network cost changes with increased protection levels to help future operators in their decision on a feasible level of resilience provisioning. Figure 4 shows a virtual network setup cost comparison for different levels of protection with PIP-Resilience and VNO-Resilience. Protection against single link/node failures and subnetwork failures is realized by using two link/node or subnetwork disjoint paths for the routing of services or the mapping of virtual links, respectively. In a subnetwork failure, all the links and nodes in that subnetwork are assumed to fail. For protection against double link failures, three link disjoint paths are utilized. The number of service nodes signifies the different source node locations of the services, where services with different destinations do not necessarily use the same routing. In our analysis, we define the setup cost of a virtual network as the summation of virtual link, node, and vir-

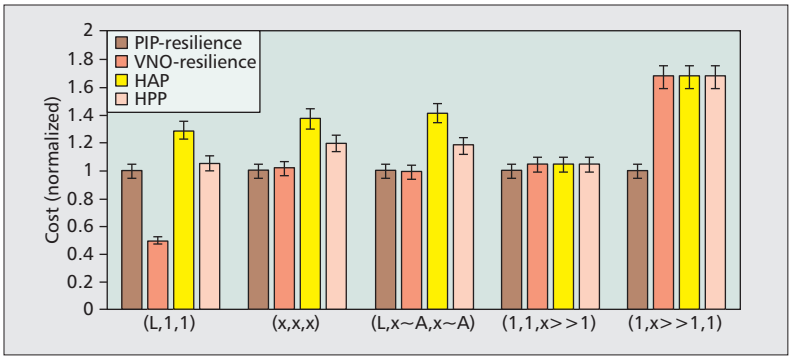


Figure 5. Virtual network cost comparison of the resilience design alternatives. The cost is defined as (virtual link cost, virtual node cost, virtual machine cost), where the setup and capacity-dependent costs of each individual resource type are equal. L is the physical length of a virtual link in kilometers, and using this option states that the cost of a virtual link is dependent on its length. A is the average shortest path length in the physical topology, and x is a positive scalar and is in the same order as A if specified as x ~ A.

tual machine (if used) costs. Each of these cost components consists of a certain fixed cost value signifying the cost of setting up this virtual element and a capacity-dependent cost value per unit capacity requested on the virtual element. For this evaluation the example cost factors are chosen such that fixed cost components are higher than the capacity-dependent cost components, and link cost is the dominant cost factor; the setup used and capacity-dependent values for the link/node cost are 200/4 and 20/4, respectively. The reason for this is the assumption that the initial setup of a virtual element can be more costly than increasing its capacity incrementally. Due to the use of fixed cost values, the relative cost behavior is mainly unaffected by the number of service nodes. Further cost analysis with varying cost values is provided in the next section using the more general case of anycast services.

Under the given assumptions, it is shown that PIP-Resilience results in a lower cost value than VNO-Resilience for all considered failure types due to the high link setup cost and higher number of virtual links required in the VNO-Resilience model. The most interesting result is that providing resilience against single link, node, or subnetwork failures has almost the same cost to an operator. In a subnetwork failure, it is assumed that all the links and nodes in a certain availability region fail simultaneously due to, for example, a disaster [11]. Thus, an intelligent virtual network design enables disaster resilience at the same cost as single link failure protection. Moreover, if protection against double link failures is requested (i.e., protection against simultaneous failure of two independent links), the cost increase compared to single link failure protection is significantly lower with PIP-Resilience than with VNO-Resilience. Finally, failures occurring in the virtual layer can only be detected and recovered from in the virtual layer. Moreover, if protection against virtual link and node failures is already provisioned in the virtual layer, it is more cost efficient to request a non-resilient network from the PIP(s) since single physical link and node protection is implicitly provided in the virtual layer.

Cost Comparison of Resilience Design Alternatives

Since it is rather difficult to estimate future cost values, the effect of varying cost parameters is analyzed to build a framework for the resilience layer decision. The cost trade-off for a VNO occurs due to the choice between renting cheaper non-resilient virtual elements but requiring a larger number of them due to redundancy provisioning, and renting a lower number of resilient higher-cost elements. The cost difference between a resilient and non-resilient resource is called the resilience premium and is taken as a multiplication factor of two in this analysis because 1:1 link and data center protection is provided. The costs are defined as tuples: (Link Cost, Node Cost, Virtual Machine Cost), as shown in Fig. 5. The cost settings are designed to show the effect of dominance or equality of the cost components. We also differentiate between fixed link cost values and link costs depending linearly on the physical length (in kilometers) of a virtual link. Fixed values are shown with a 1 (unit cost) or x, which is a real value larger than 0, and the case with length dependence is specified with an L. The fixed and unit capacity cost values of each component are assumed to be equal to simplify the comparison. The simulation results are within a ±5 percent confidence interval with a confidence level of 95 percent. The results are shown for two randomly located data centers owned by a single PIP and 10 service source nodes, where a single virtual network solution can be computed within a few seconds on a computer with 16 cores and 60 Gbytes RAM memory. The results are scaled down to cost = 1 for PIP-Resilience for each case to allow comparison of the models with the different cases, but each alternative has different absolute values and can use different service routing and virtual resource mapping.

For (L,1,1), where virtual link cost is dependent on the physical length of the link, VNO-Resilience results in a virtual network cost value lower than half those with other resilience alternatives due to its routing advantage compared to PIP-Resilience and the usage of disjoint virtual paths containing virtual links using simple path mapping compared to the hybrid models. Having equal emphasis on all cost components, as in (x,x,x) and (L,x ~ A,x ~ A), causes VNO-Resilience and PIP-Resilience to perform very close to each other and better than both hybrid alternatives as node cost compensates the routing advantage of VNO-Resilience and increases the cost of hybrid models. If the virtual machine or node cost is the dominant cost component as in (1,1,x >> 1) and (1,x >> 1,1), VNO-Resilience, HAP, and HPP result in almost equal values, and PIP-Resilience has a lower cost due to its lowest virtual node resource requirements. If virtual machines dominate the cost, the cost with PIP-Resilience is only slightly better, but with

dominance of node cost the difference is significant. The results in Fig. 5 are observed for a single data center provider, where increasing the number of the data center providers, the distance between individual data centers, and the number of service nodes makes VNO-Resilience more favorable than PIP-Resilience, and reduces the excess cost in HAP and HPP for length-dependent virtual link cost. For the other three cases, the results remain in the same range.

In conclusion, the cost performance of resilience designs depends heavily on the actual cost values. PIP-Resilience is favorable if the node cost is dominant. With a dominant link cost, VNO-Resilience performs the best. For equal cost values, having resilience entirely in either the virtual or physical layer is a better option than hybrid designs. Where virtual machine cost dominates in terms of virtual network cost, the operator is rather free to decide on the layer of resilience provisioning. In such a case, other criteria to consider can be the required failure coverage and the level of network management knowledge at the VNO layer.

Conclusion

This article tackles the question of how to provide end-to-end resilience for cloud services in case of failures and disasters, and proposes a solution based on network virtualization. After the introduction of a detailed architecture and resilient virtual network design solutions, we investigate at which layer to provision resilience in terms of failure coverage and virtual network setup cost — the fee a VNO needs to pay to PIPs for rental of virtual resources and establishment of a virtual network. With the used cost model, we show that providing resilience against single link, node, or subnetwork failures have almost the same cost to a VNO and a PIP. Failures occurring in the virtual layer can only be detected and recovered from in the virtual layer. If protection against these failures is already in place, delegating protection against physical link and node failures to PIPs is not needed, since it is implicitly provided. We also provide a detailed analysis from the cost perspective with various pricing alternatives offering a framework in the decision on realizing resilience in the virtual or physical layer, or a combination of both. Future work may address analyzing the resilience layer in terms of, say, service latency, resource requirements, and complexity. Moreover, the effect of dynamic server resource allocation and redundant capacity sharing is a topic for further evaluation. Finally, the models' protection level can be adjusted, for example, by only protecting the primary path for PIP-Resilience, and the effect of such adjustments should be investigated.

References

[1] Symantec, Virtualization and Evolution to the Cloud Survey, 2011.
[2] The Year in Downtime: Top 10 Outages of 2012 and 2013, http://www.data centerknowledge.com, 2012-2013.
[3] G. Schaffrath *et al.*, "Network Virtualization Architecture: Proposal and Initial Prototype," *1st ACM Wksp. Virtualized Infrastructure Systems and Architectures*, NY, 2009
[4] M. Hoffmann and M. Staufer, "Network Virtualization for Future Mobile Networks: General Architecture and Applications," *IEEE ICC Wksps. AMN 2011*, June 2011.
[5] L. Fang *et al.*, "BGP/MPLS IP VPN Data Center Interconnect," Internet draft (work in progress), IETF, Oct. 2013, draft-fang-l3vpn-data-center-interconnect-02.
[6] D. Anderson *et al.*, "Resilient Overlay Networks," *Proc. 18th ACM Symp. Operating Systems Principles*, 2001.
[7] D. Dietrich, A. Rizk, and P. Papadimitriou, "Multi-Domain Virtual Network Embedding with Limited Information Disclosure," *IFIP Networking Conf. 2013*, May 2013.
[8] I. B. Barla *et al.*, "Optimal Design of Virtual Networks for Resilient Cloud Services," *9th Int'l. Conf. Design of Reliable Commun. Networks*, Budapest, Hungary, 2013.
[9] I. B. Barla *et al.*, "Shared Protection in Virtual Networks," *IEEE ICC '13 Wksp. Clouds, Networks and Data Centers*, Budapest, Hungary, 2013.
[10] C. Develder *et al.*, "Survivable Optical Grid Dimensioning: Anycast Routing with Server and Network Failure Protection," *IEEE ICC*, June 2011.
[11] A. Basta *et al.*, "Failure Coverage in Optimal Virtual Networks," *OFC/NFOEC*, 2013, paper OTh3E.2.

Biographies

Isil Burcu Barla Harter (barla@net.in.tum.de) is a Ph.D. candidate at Technische Universität München (TUM) and Nokia Munich, Germany. She received her Bachelor's degree from Bogazici University, Istanbul, Turkey, in 2007, and her Master's degree from TUM in 2009. Her research interests include network virtualization, cloudification, recovery methods, network optimization, routing, and security.

Dominic Schupke (dominic.schupke@airbus.com) is with the Airbus Group (previously EADS) in Munich, Germany, working in the Innovations unit in the area of wireless communication. Prior to EADS he was with NSN, Siemens, and TUM. He received his diploma from RWTH Aachen in 1998 and his Ph.D. degree from TUM in 2004. Since April 2009 he has taught the Network Planning course at TUM. His research interests include network architectures and protocols, routing, recovery methods, availability analysis, critical infrastructures, security, virtualization, network optimization, and network planning.

Marco Hoffmann (marco.hoffmann@nsn.com) studied computer science and received the Dr rer. nat. degree from TUM in 2005. In 2004 he joined the Research and Development Department of Siemens. Currently he is Technology Manager and Project Manager for international projects in the Research division of Nokia. He has been a consortium leader and board member of several national and international projects and member of company internal and nation-wide future Internet strategy teams.

Georg Carle (carle@in.tum.de) is a professor at the Department of Informatics of TUM, holding the chair for Network Architectures and Services. He studied at the University of Stuttgart, Brunel University, London, and Ecole Nationale Superieure des Telecommunications, Paris. He received his Ph.D. in computer science from the University of Karlsruhe, and worked as a postdoctoral scientist at Institut Eurecom, Sophia Antipolis, France, at the Fraunhofer Institute for Open Communication Systems, Berlin, and as a professor at the University of Tübingen.

Future work may address analyzing the resilience layer in terms of, say, service latency, resource utilization, and complexity. Moreover, the effect of dynamic server resource allocation and redundant capacity sharing is a topic for further evaluation.

COMMUNICATIONS EDUCATION AND TRAINING: EXPANDING THE STUDENT EXPERIENCE



David G. Michelson *Maria Trocan* *Wen Tong*

This is the second Feature Topic on Communications Education and Training; these will appear regularly in *IEEE Communications Magazine* henceforth. It has long been appreciated that successful design, implementation, maintenance, optimization, and improvement of modern communications networks require the efforts of a highly educated, well trained, and dedicated workforce. In recent years, various initiatives have been undertaken in both academia and industry that seek to improve our capacity to educate and train both current and future communications workers. These range from new learning technologies and pedagogies and new university accreditation programs to advanced cooperative education programs, training alliances, and industry certification programs.

The ComSoc Education and Training Board is sponsoring this Feature Topic in order to promote sharing of recent efforts to advance the state of the art in a discipline that is so vital to the long-term health of both our field and the careers of our members. Our ultimate goal is to hasten the adoption of promising new methods and techniques by our community. The focus of this month's Ffeature Topic is on the manner in which universities are expanding the student experience by adopting novel approaches to communications education.

The limitations of the traditional lecture and laboratory in delivering engineering knowledge and insights to students have long been recognized. Over the past 30 years, a variety of innovative techniques have been developed to more fully engage students. These include team-based learning, problem-based learning, project-based learning, and community-based learning. While instructors and students have found merit in these approaches, implementing such schemes places nontrivial demands on both time and resources. Sharing best practices is vital if such approaches are to find broad success and acceptance.

The two articles presented in this section provide straightforward descriptions of their authors' respective experience in delivering innovative approaches to communications education along with reflective insights concerning the experiences of both instructors and students. Anyone wishing to establish their own innovative curricula will benefit from the stories told here.

In "A Project-Oriented Learning Experience for Teaching Electronics Fundamentals," Frédéric Amiel, Dieudonné Abboud, and Maria Trocan (Institut Supérieur d'Electronique de Paris) describe an experiment conducted over a six-year period at ISEP based on a project-oriented learning approach to teach the fundamentals of electronics. The proposed teaching framework has a dynamic structure, as it adapts and modifies the terms of annual assessments to foster motivation and interest of students. The results show the effectiveness and value of this approach for student motivation.

In "Bringing an Engineering Lab into Social Sciences: A Didactic Approach and an eExperiential Evaluation," Jesus Cano, Roberto Hernandez, and Salvador Ros (Universidad Nacional de Educación a Distancia) explore the challenge and experience of developing skills that are typical of information and communications engineering but are linked to cyber security in an unusual academic context within the branch of social sciences, specifically for law and criminology students. They describe how use of game-based learning has proven to be an exceptionally effective method of crossing this gap.

The ComSoc Education and Training Board is committed to promoting the sharing of such experiences through this Feature Topic. Please consider sharing your own experience in response to future calls. Our next Feature Topic is concerned with industry certification and university accreditation programs. Submissions are due on 15 January 2015. The Feature Topic will be published in the May 2015 issue.

BIOGRAPHIES

DAVID G. MICHELSON [S'80, M'89, SM'99] (davem@ece.ubc.ca) received his B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from the University of British Columbia (UBC), Vancouver, Canada. From 1996 to 2001, he served as a member of a joint team from AT&T Wireless Services, Redmond, Washington, and AT&T Labs-Research, Red Bank, New Jersey, where

he contributed to the development of propagation and channel models for next-generation fixed and mobile wireless systems that formed the basis for those later adopted by the WiMAX and LTE communities. Since 2003, he has led the Radio Science Laboratory in the Department of Electrical and Computer Engineering at UBC, where his research interests include antenna design and channel modeling for wireless sensor networks, advanced cellular, short range vehicular, and satellite communications. In 2011, he and his former student Simon Chiu won the 2011 R. W. P. King Best Paper Award of the IEEE Antennas and Propagation Society. He is a member of the Boards of Governors of the IEEE Communications Society and IEEE Vehicular Technology Society, past Director of Education and Training for ComSoc, and serves as General Chair of the IEEE Vehicular Technology Conference 2014-Fall, the 2015 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting, the 2015 International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities (TRIDENTCOM), the 2015 IEEE International Symposium on Ethics in Engineering, Science, and Technology (IEEE ETHICS), and is the founding Editor of the Wiley/IEEE Press Series on Vehicular Technology.

MARIA TROCAN [SM] (maria.trocan@isep.fr) [M'09, SM'13] received her M.Eng. in electrical engineering and computer science from Politehnica University of Bucharest in 2004, her Ph.D. in signal and image processing from Telecom ParisTech in 2007, and her Habilitation to lead research (French Habilitation à Diriger des recherches) in 2014. She joined Joost, Netherlands in 2007, where she worked as a research engineer involved in the design and development of video transcoding systems. Since May 2009 she has been an associate professor with the Signal, Image and Telecommunications Department at Institut Superieur d'Electronique de Paris (ISEP), where since October 2011 she has been responsible for the Signal Processing Graduate Program. She is Associate Editor for the Springer *Journal on Signal, Image and Video Processing*. In 2014 she was elected IEEE France Vice-President responsible for Student Activities, and an IEEE Circuits and Systems Board of Governors member, responsible for the Young Professionals Program. She is currently serving as a reviewer for major IEEE journals and conferences, and has been involved in various conference organization and program committees.

WEN TONG [S'89, F'14] (tongwen@huawei.com) is the head of Wireless Research, the Communications Technologies Laboratories, and the Huawei 2012 Lab, and is a Huawei Fellow. Prior to joining Huawei in March 2009, he was Global Head of the Network Technology Labs at Nortel and a Nortel Fellow. He received M.Sc. and Ph.D. degrees in electrical engineering in 1986 and 1993, and joined the Wireless Technology Labs at Bell Northern Research in Canada in 1995. He has pioneered fundamental technologies in wireless with 210 granted U.S. patents. He was Nortel's Most Prolific Inventor. He has conducted advanced research work spanning from 1G to 4G wireless at Nortel. He had been the director of Wireless Technology Labs from 2005 to 2007. From 2007 to 2009, He was the head of Network Technology Labs, responsible for Nortel's global strategic technologies research and development. In 2007, he was inducted as a Nortel Fellow. Since 2010, he has been the vice president and head of Huawei Wireless Research, leading one of the largest wireless research organizations in the industry with more than 700 research experts. In 2011, he was appointed head of the Communications Technologies Labs of Huawei 2012 LAB, a corporative centralized next generation research initiative. In 2011, he was elected as a Huawei Fellow. He serves on the Boards of Directors of the WiFi Alliance and the Green Touch Consortium.

A Project Oriented Learning Experience for Teaching Electronics Fundamentals

Frédéric Amiel, Dieudonné Abboud, and Maria Trocan

ABSTRACT

This article describes an experiment conducted during six years at the Institut Supérieur d'Electronique de Paris based on a Project Oriented Learning approach to teach the fundamentals of Electronics. The proposed teaching framework has a dynamic structure, as it adapts and modifies the terms of annual assessments to foster the motivation and interest of students. Results show the effectiveness and the value of this approach for student motivation.

INTRODUCTION

Classical pedagogy, based on the transmission of knowledge by the famous course-seminaries-lab work scheme, has been proven more and more inadequate for higher education (e.g. bachelor and master degrees), especially when it comes to vocationally oriented studies.

The demotivation phenomenon has also been encountered at the Institut Supérieur d'Electronique de Paris (ISEP), an Information Technology Engineering School, in an amplified manner due to the saturation effect shown by the engineering students concerning the theoretical approach, which is a characteristic of the engineering education model in France.

Indeed, after two years of preparatory classes focused on basic science (26 hours per week of mathematics and physics) [1] with an emphasis on abstraction, modeling, and computational capacity, the students who reach the engineering education level are low motivated by these theoretical developments. This phenomenon is especially encountered in the field of modern physics applied to Telecommunications, Signal Processing, and Electronics. However, we see these same students strongly engaged in applicative lessons or technical projects associated with the classes in these rather difficult areas.

In fact, the motivation behavior of the students at ISEP is similar to current digital natives [2]. They adhere to instruction (contents and methods) if they understand a genuine interest in relation to their training project. In other words, it is the final purpose of the course that is important to them, and the saturation effect mentioned above does sharpen their sensitivity to the needs of the professional world.

This fact cannot be ignored and motivates adaptive and innovative efforts. Thus, a profound reflection was completed at ISEP concerning the teaching methods, the contents of training programs, and their purposes.

Simultaneously, innovative experiments have been undertaken to improve student motivation. We were looking primarily to link the theoretical courses in mathematics and physics with targeted technological breakthroughs to bring together the concepts and abstract structures of their applicative context.

We tried thematic studies and projects with a professional nature performed in small groups [3]. The encouraging results of these experiments and the analysis of several educational innovations geared toward active learning [3] led us to choose:

- A competency-based approach, explicitly identifying the purposes of training and serving as a structural element of the program content.
- Project oriented teaching, in order to develop a set of hard and soft skills that meet the real needs of the professional world.

This article describes the Project Oriented Learning (POL) in Electronics experimented at ISEP during six years. We will describe the proposed project and its organization in the second section. Student evaluation and some measurements concerning the motivation are proposed in the third section. The final section draws the conclusions of this experiment.

PROJECT DESCRIPTION

CONTEXT

The training project at ISEP includes fundamentals of Electronics for all the students during the first engineering year. The class curriculum includes the following:

- Amplification, filtering, recovery, and conditioning of signals from sensors.
- Fourier series and transforms in the context of signal processing.
- Analog filtering and basics of digital filtering.
- Analog to digital conversions.
- Basic schemes of data transmission (coding, channel concept, communication parameters).
- Combinatorial and sequential logics.

The authors are with Institut Supérieur d'Electronique de Paris.

0163-6804/14/$25.00 © 2014 IEEE

Some more advanced groups will address:
- Power switching.
- Computer interfacing.
- Usage of more advanced communication components.

Regarding the know-how, students will:
- Program an FPGA.
- Implement and perform electrical measurements.
- Use simulators and modeling tools.

The V-model is used for the project management to teach the top-down system approach. The students must learn how to specify and propose a system using some modules for which only the input/output specifications are given (black boxes).

In France the engineering curriculum is organized as follows: after completing the baccalaureate, the students follow two years of intensive preparatory courses, where they learn the fundamentals of mathematics and physics. Further, they pass some contests to join a three-year engineering school in a vocationally oriented cycle.

In terms of a technological teaching program, the first year at ISEP proposes the fundamentals of Information Technology (programming, use of databases, and some web technologies), introduction to computer architecture and the proposed project for teaching the fundamentals of Electronics.

After the global training overview in the first year, students must choose a vocational oriented teaching track in one of four fields — IT, Electronics, Signal Processing, and Telecommunications — and receive more professionally dedicated information. They will also achieve 10 months of work experience and complete a semester in a university abroad.

In the first engineering year, the students' knowledge in Electronics comes from the preparatory school's physics course. They know the basic laws of electricity and also have conducted some electrical measurements experiments.

Project Oriented Learning Description and Conditions

Two days a week during a semester (i.e. 15 weeks), the students must pursue the following objective:
- Create a small autonomous trolley, capable of moving and looking for a tag.
- As the tag sends some data, the system should be able to decode, process the received information, and send a feedback message.

This general topic varies each year (e.g. communication with peers, sound, light detection, various messages processing etc.). Teams of six students work in a small room with a round table and a workbench. They receive a motorized base and an FPGA board connected with a printed circuit, permitting them to implement analog functions. Two computers with the needed software and a good-sized WiFi network are also provided. Three workshops accelerate the handling of the needed software (FPGA programming, VHDL concept, and the electrical simulator).

Each team realizes the project while following a booknote, describing the tasks and the goals to achieve. A supervisor follows five groups during two half-days per week (hence half-time of the student work), in order to provide a methodological and technical assistance. The original aspects of the proposed POL are:
- A longer period than conventional Project Oriented Learning [4, 5].
- Consequently, a more complex and transdisciplinary project [4, 5].

Evaluation/Development

Evaluation Method

Different methods have been tested in order to assess the competency level of students and to ensure that those who successfully complete the module have reached the required level.

Supervisor Evaluation: Classical exams that existed during the first four years of project implementation to test the knowledge level of students have been removed for the last two years, the evaluation being made by the supervisor using a multi-criteria grid. This evaluation is done through the complete project duration and each student can access any time his evaluation and see his own progress.

Approximately every three weeks the students' team must present its work to the supervisor, who is assisted sometimes by an external teacher, during a formal presentation. This exercise is followed by a question and answer session that helps the supervisor formalize the evaluations and to discuss them with each student.

Cross-Evaluation: The students are required to cross evaluate themselves after two months and also at the end of the project. This ensures that the supervisor understands the group's dynamics and has detected the leaders, the students who work the most, and those who work the least.

Final Presentation: At the end of the project each team has to provide a report and present its work in front of a two-teacher jury.

White Paper Project: After the project's completion, the weakest students (following the supervisor evaluation) will pass an extra examination to verify if they can translate their skills to another subject.

All of these results are evaluated by the supervisor in order to determine who achieved the required competency level.

Project Evolution

Another critical point is to verify the student's motivation as the Project Oriented Learning scheme was chosen in order to increase the interest of the student for studying Electronics. We proposed an annual quiz to obtain students' impressions of this Project Oriented Learning approach. Each year we discover the necessity to adapt and improve our method by different ways in order to follow the students' versatility.

Transition — We discovered during the first two years of the POL experiment that students were confused and hardly felt the interest of this method compared to the conventional way of teaching. Teamwork was felt unfair and working on a project seemed too far from the theory to understand it. This learning model followed a bottom up approach in contrast to the classical top (theory)

We proposed an annual quiz to obtain students' impressions of this Project Oriented Learning approach. Each year we discover the necessity to adapt and improve our method by different ways in order to follow the students' versatility.

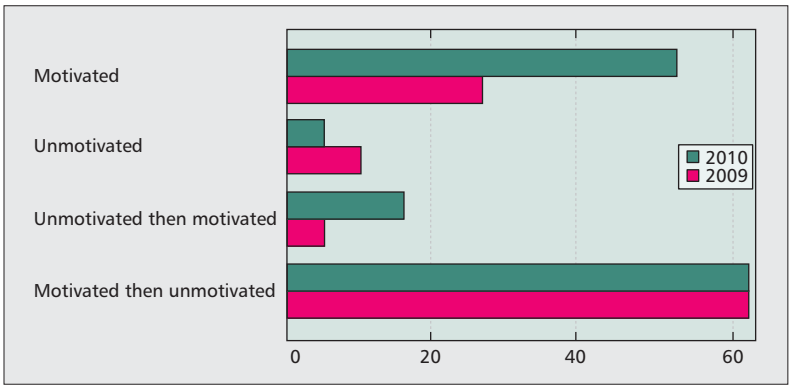


Figure 1. Motivation trend during the semester — two year comparison.

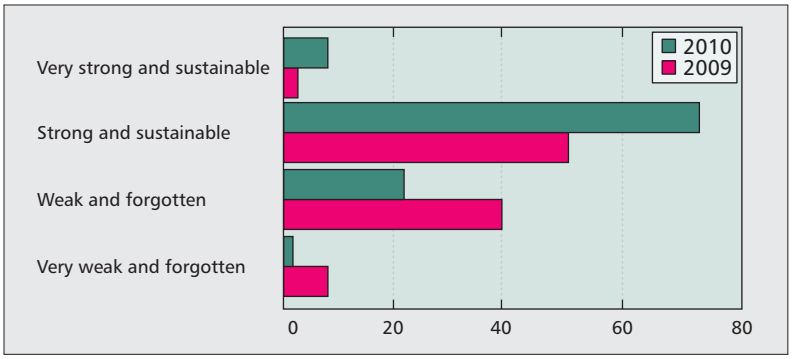


Figure 2. Students impression on their acquired competency stability two year comparison.

down (applications) approach, to which they were accustomed. They also felt that the competency approach is rough and hard to understand.

Consequently, in order to smooth this transition we have scheduled, at the beginning of their engineering studies and during three weeks, a pedagogical break. In this period the students assist with presentations from alumni who are engineers in different professional environments. They will also visit some heterogeneous oriented companies, in the professional tracks provided by ISEP. This approach proved efficient: at the end of this period, 80 percent of the students feel curious and motivated by the POL they will attend.

Motivation [5] — Student motivation during the five month project is a challenging issue, as can be seen in Fig. 1. We discovered the importance of the following points:

- Logistical problems must be handled efficiently.
- Bad-time repartition: during the first four-year examinations, the students passed time to appropriate the information just before the evaluation day.
- The required autonomy, needed by the POL approach, confuses some of the students.

We handled all these points by removing the examinations and by preventing and training the supervisor. The effect of this policy is shown in Fig. 1.

Team Work — As the project represents a half semester of work, the students feel very motivated to succeed. In this context, the investment level and the academic level disparity of the students inside each team is a source of contention [6]. The supervisor has to be very aware of the team ambiance and explain carefully his evaluation. At the project's end, 60 percent of the students feel "satisfied" or "very satisfied" by working in a team.

Knowledge Stability — Finally, we ask the students to express their opinions concerning the stability of the acquired competencies. Fig. 2 shows that the more we support and prepare the ground in the students' mind, the higher is the effectiveness of the training.

Conclusion

A project-based learning approach to teach the fundamentals of Electronics has been proposed as an alternative to classical course-lab training, to better cope with current professional requirements. The effectiveness of such a teaching methodology resides in the establishment of a constant feedback mechanism.

References

[1] www.education.gouv.fr
[2] C. Jones and B. Shao, *The Net Generation and Digital Natives: Implications for Higher Education*, Higher Education Academy, York, UK, 2011.
[3] L. Pereles, J. Lockyer, and H. Fidler, "Permanent Small Groups: Group Dynamics, Learning, and Change," *The Journal of Continuing Education in the Health Professions*, 2002.
[4] L. Arpin and L. Capra, *L'Apprentissage Par Projets*, Chenelire/McGraw-Hill, Canada, 2001.
[5] T. Markham, "Project Based Learning," *Teacher Librarian*, vol. 39, no. 2, 2011, pp. 38–42.
[6] C. Ames, "Achievement Goals and the Classroom Climate: Students' Learning Strategies and Motivation Processes," L. Erlbaum (dir.), *Student Perceptions in the Classroom*, Hillsdale: L. Erlbaum, 1992, p. 327-208.

Biographies

Frédéric Amiel obtained his DEA (MSc) in information systems in 1985, one year after his electronics engineering degree. He worked in several companies as a hardware developer and joined ISEP in 1992. Since then he has been an associate professor responsible of the digital electronics laboratory. He has also worked as a consultant engineer for several companies and was in charge of several teaching programs in digital electronics, DSP, and FPGA technologies. He is currently in charge of the electronics program and embedded systems major at ISEP.

Dieudonné Abboud received his Ph.D. in physics didactics from Paris 7 — Jussieu University in 1984. After his Ph.D. he worked for the Lebanese Education Ministry, where he was responsible for the Pedagogical Orientation Unit. In 1987 he joined Pierre and Marie Curie University in France as a researcher in physics didactics. Between 1994 and 2002 he was the dean of studies with the Institut Supérieur d'Ingénierie Appliquée in Paris. Since 2003 he has been the dean of studies with the Institut Supérieur d'Electronique de Paris.

Maria Trocan received her B.Eng. in electrical engineering and computer science from Politechnica University of Bucharest in 2004, her Ph.D. in signal processing from Telecom ParisTech (formerly Ecole Nationale Supérieure des Télécommunications) in 2007, and the habilitation to lead researches ("Habilitation Diriger des Recherches") from Pierre and Marie Curie University in 2014. She joined Joost–Netherlands in 2007, where she worked as a research engineer involved in the design and development of video transcoding systems. Since May 2009 she has been an associate professor with the Signal, Image and Telecommunications Department at the Institut Supérieur d'Electronique de Paris (ISEP).

Bringing an Engineering Lab into Social Sciences: Didactic Approach and an Experiential Evaluation

Jesús Cano, Roberto Hernández, and Salvador Ros

ABSTRACT

In this work we explore the challenge and experience of developing skills that are typical of information and communications engineering but linked to cybersecurity in an unusual academic context within the branch of social sciences, specifically for law and criminology students. We examine the prior assumptions regarding the technical issues evaluated, the result of putting the designed laboratory into practice, and also student satisfaction with the experience. The engineering skills-based experiences should be incorporated transversally into other subjects of non-technological sciences. Our research contributes experience along these lines, and inspires confidence in this respect for the new and ever more digitally native batches of students.

INTRODUCTION

The concern for training in cybersecurity at all levels is of interest in the academic world. The interaction between engineers and non-technical professionals is currently a major focus of education in this matter [1]. At all levels, but particularly in the university context, it is deemed necessary to update and profit from the experience and know-how of private enterprise and government institutions [2]. Professional bodies such as the Association of Computer Manufacturers (ACM) or IEEE are working in this direction in the Joint Task Force on Computing Curricula in order to produce a reference syllabus of computing sciences [3].

Our research presents the analysis, design, and experience of a course in cybersecurity for criminology. To take on this task, an interesting educational approach could be to use game-based learning (GBL), an emerging trend that can enable us to comfortably cross the bridge between engineering and social sciences.

GBL, particularly serious games, are currently a clearly emerging trend in various fields, such as education, technology, and vocational training. The IEEE predicts that in 2020 gaming will be incorporated into more than 85 percent of daily tasks, with the result that game elements will be a standard part of the way things are run in both business and education [4].

Some research is coming to explore the state of the art of security-related games so that they may be of use for engineering studies, where some games oriented to the teaching of engineering and network management are specifically examined. Other authors have contributed games, with CyberCIEGE, a simulator in video game format to simulate network security environments and mechanisms; it is an outstanding game in the field of cybersecurity [5].

Several models have been developed that identify distinct learning outcomes that playing digital games can have. Thus, in [6] the authors have identified the cognitive skills regarding games and have grouped them into five "families of cognitive demands," which are understanding content, problem solving, teamwork, communication, and self-regulation. The work proposed in [7] considers learning outcomes to analyze and emphasize the importance of instructional support in game-based learning through critical review of the literature.

Our research presents a serious games-based cybersecurity course design for teaching the subject "Security and Emergency Systems and Plans" (12 ECTS) within the area of Information Security, part of the Criminology and Security Sciences degree, an undergraduate course taught at the Faculty of Law of the San Pablo CEU University in Madrid, Spain.

This work is organized as follows. First, the methodology used to build a games laboratory is described. We then set out how the games laboratory was built for the specific subject, taking into account the foregoing criteria and evaluating the outcomes. The final section sets out the conclusions.

METHODS

Since the aim of the work is simple to : — to design a games-based cybersecurity course for criminologists — the methods for doing so should detail the building process in sufficient detail for it to be reproducible. This section analyzes this process, the criteria adopted, and the decisions made by the teaching team in each part of the process.

The authors are with the Spanish University for Distance Education (UNED).

0163-6804/14/$25.00 © 2014 IEEE

Knowledge areas study the mechanisms of prevention, investigation techniques, treating the victim, and why, how, and when the criminal act takes place. This means there is not only a heavy load of penal law, psychology, and sociology but also a scientific, medical, and technological load.

1	Principles of security
2	Policy, plans, and procedures
3	The human factor
4	Vulnerabilities, threats, and malware
5	Cyberterrorism, cyber espionage, and criminal organizations
6	Incident response and computer forensics
7	Access to information and applied cryptography
8	Network security scenarios
9	Wi-Fi and ubiquitous crime
10	Forgery, scams, and phishing
11	Phenomenology of electronic payment and economic crime
12	Computer crime
13	Personal data protection

Table 1. Topics of cybersecurity in the training plan.

A Study of the Educational Context

Nowadays, the link between criminology and cybersecurity is obvious. With the advances in technology and its generalized use in current society and the global context, deviant behavior is also digitally instrumentalized. Criminology is a security science that studies man's criminal effects and antisocial behavior. Knowledge areas study the mechanisms of prevention, investigation techniques, treating the victim, and why, how, and when a criminal act takes place. This means there is not only a heavy load of penal law, psychology, and sociology, but also a scientific, medical, and technological load.

On the other hand, the common corpus of cybersecurity includes basic disciplines such as computer and engineering sciences, law, ethics and psychology, business management and teaching, and sociology and its criminal aspects.

Our fieldwork was set up for third year students in the criminology degree in a computer network room with Internet connection. There were 32 student volunteers in all, 17 men and 15 women, for one academic year. Therefore, the male-female student profile was relatively balanced, being approximately 50 percent for the course studied. The group covered a target age of 20–24 years old. It should be pointed out that a control group could not be set up since all the students wanted to take part in the laboratories and thought it "unfair" if they could not do so.

Table 1 shows a general outline of topics in the training plan that has been designed. The subject has 12 European Credit Transfer System (ECTS) credits that are distributed among the areas of physical security, electronic security, and information security. In this work, we concentrate on the latter area. The ratio of theory classes to laboratory practice was 2:1; that is, two hours of theoretical content in the classroom for each hour of work in the laboratory. It should be borne in mind that the software used in the laboratory is freely distributed, and students can work with it anywhere, not only in the laboratory. However, the hours voluntarily devoted by students have not been included in the description.

Research Stage of the Prior Assumptions

Taking the university context as a starting point, a study was made of students' assumptions regarding criminology, their prejudices in the subject, and their fears. The multidisciplinary nature of the subject and the qualification must be understood as a challenge, and the difficulties entailed must be overcome.

The survey technique was used with brief responses and/or identification of concepts. A model was designed to evaluate the assumptions based on 20 main items and 26 cybersecurity descriptors. The aim of the survey was to establish an initial perspective of the teaching-learning process. In this way, conceptual and attitudinal content, and procedural predisposition could be evaluated by following the set training plan. Table 2 shows these items and the descriptors.

Evaluation item i3 is broken down into a set of descriptors that explore conceptual aspects related to cybersecurity. Twenty-six descriptors have been established, as Table 2 shows. The purpose of this model is to mark and identify the concepts that are familiar to students.

Prior Evaluation — The empirical result of the prior assumption items survey is in the form of a double-entry report. This is the result of the analysis of the responses and their classification.

Figure 1 lets us see the view of the prior assumptions that results in a negative impression due to the difficulty, ignorance of, or aversion to the item evaluated. The most marked are items i8, i9, and i15. The negative ideas are linked to the organization of security, personal experience of incidents, and the sensation of risk.

As for the positive ideas evaluated, items i1, i3, i4, i5, i10, i12, and i20 show positive results. Parameters i1, i4, and i5 evaluate global aspects that reveal a positive predisposition for approaching cybersecurity learning. The positive attitude shown regarding collaboration on the Internet, especially the use of social networks (items i10 and i12), is interesting. Item i20 is a meta-view of the study of prior assumptions itself, which, as we can see, has been accepted as a positive initiative. Item i3 evaluates the conceptual content of the set of descriptors in Table 2 using an identification and marking procedure, the results of which are dealt with further on. Figure 1b shows these results as a graph too.

The results regarding security terms let us define an initial profile for organizing the teaching-learning process. Particularly marked are descriptors a, b, h, j, v, and z, which must be the starting point for the framework of the laboratory theoretical work. There is an initial conceptual identification of the terms referring to cybersecurity, cyber threats, trojans, viruses,

cookies, and personal data. These can form the cognitive basis of our laboratory. The collection of descriptors f, k, p, q, r, u, and x, and with a low score explore the terms script kiddies, pentest, skimming, stenography, continuity plan, IDS, and DMZ. In this respect, we can discern a variety of gaps in security protection and organization structures, which need to be strengthened by designing the laboratory process.

Preliminary Discussion— Consequently, the initial attitude and predisposition of criminology students is good since it can be deduced that the general perception shown toward the subject is more positive than negative. As a consequence, we have studied the positive aspects of their feelings on one hand and the negative ones on the other. Note that by superimposing the areas evaluated in Fig. 1, the calculation is positive.

The starting point for evaluating the descriptors that were defined for the conceptual content has now been established as a general expression of the overall evaluation. Statistically, the modal value of the study of this qualitative distribution is zero, which gives a figure of 27 percent for totally unknown concepts. This enables us to have a rating which can be used to form a learning strategy with content based on concepts that are meaningful to students.

Research Stage for Developing Lab

Taking into account the nature of the course to be taught and the analysis carried out on the prior assumptions of the students taking the course, serious games emerge as a useful tool for facilitating the acquisition of knowledge.

The purpose of the games in the course is to facilitate study, not to avoid it. They are a useful supplement for helping the understanding of *processes*. Also, insofar as possible, they should lead to an increase in learning productivity and enable more technically complex content that would normally fall outside the aims of the course to be approached but which, conceptually, can be studied to explain events and how they occur. The experience of "living" the process helps understand the facts and facilitates communication between professionals with a technological background and those with a socio-judicial background.

Methods — This section describes the methods contemplated in the research to design a laboratory as a teaching resource to aid the cybersecurity training set as part of the criminology course.

The course design was approached by basing it on learning theories and guided instructional design so that students could construct their own experience through short games with specific guidelines commented on by the teacher.

The laboratory methodology comprises three parts. First, each game is presented with an introduction on how it is related to the course aims, and a motivating summary of how the game addresses their personal and professional concerns. Second, the students get together in pairs, which facilitates group work, the exchange of opinions, the opening up of group debates, information searches, and exploration using the Internet, thereby leading to active participation. Finally, each student must undertake an individual self-assessment based on the results of each game and express their impressions of satisfaction regarding each learning-game unit.

#	Item to be evaluated
I1	Survey and general expectations
I2	Concept of information
I3	**Key cybersecurity descriptors:** a. Cybersecurity b. Cyber threats c. Computer risk d. Vulnerability e. Exploits f. Script kiddie g. Worm h. Trojan i. ISO 27001 j. Virus k. Pentest l. Cryptology m. Spyware n. Cryptosystem o. Scamming p. Skimming q. Stenography r. Continuity plan s. Contingency plan t. Phishing u. DoS v. Cookie w. Firewall x. IDS y. DMZ z. Personal data
I4	Initial attitude to the subject
I5	Concerns and preferences
I6	Motivation for cybersecurity
I7	Sociological and moral aspects
I8	Organising security
I9	Personal experience of incidents
I10	General collaboration in the Internet
I11	Ideas on email abuse
I12	Survey on social networks
I13	Moral position about emerging technologies
I14	Ubiquitous-pervasive computing dilemmas
I15	Sensation of risk
I16	Psychology and prejudices
I17	Ideas about cryptography processes
I18	Concept of digital signature
I19	Open survey
I20	Meta-evaluation and predisposition

Table 2. Prior assumptions model for the survey on cybersecurity.

The educative process takes into account two complementary pedagogical models. Gagné's nine events of instruction model was considered when choosing the games, with the purpose of deciding the games' capability to direct learning toward the desired outcomes from a functional point of view [8], and Keller's ARCS model for taking motivational components into account [9].

Finally, we believe that games should be a

Taking the university context as a starting point, a study was made of students' assumptions regarding criminology, their prejudices in the subject, and their fears. The multidisciplinary nature of the subject and the qualification must be understood as a challenge, and the difficulties entailed must be overcome.

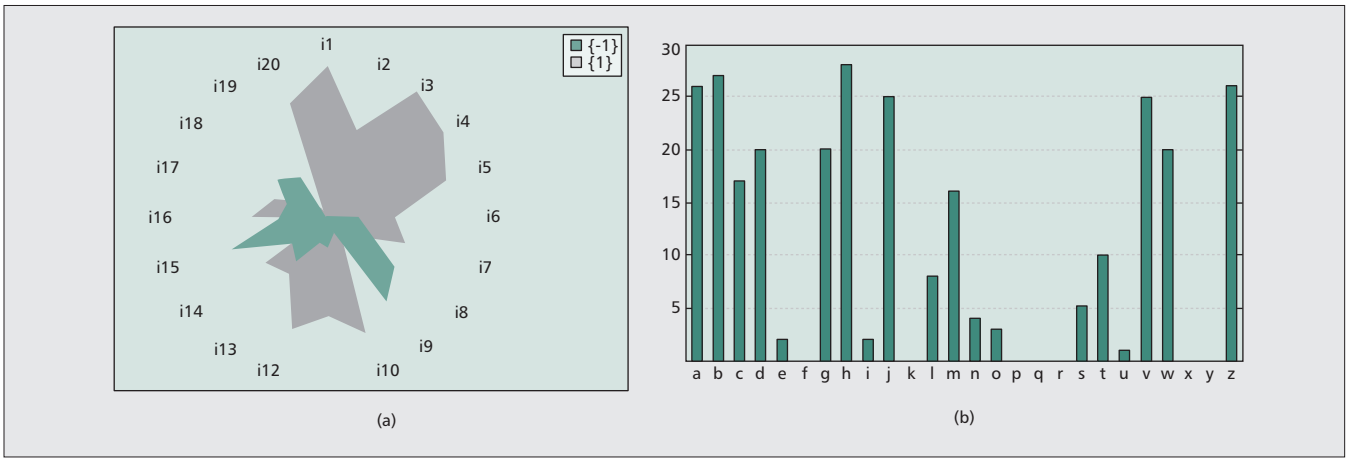


Figure 1. Representations of prior evaluation: a) positivity vs. negativity; b) descriptor data.

tool for not only facilitating self-learning throughout the course but also future self-learning. Therefore, the types of games to be used must be decided before choosing specific games.

Therefore, the following selection criteria can be set:

- The types of games to be used
- The subject content covered by the game
- The instructional value of the game
- The motivational value of the game
- The added value contributed

The first decision to be made is what games to include. An initial alternative approach is to use a known product. CyberCIEGE [5], a well-known tool, could be used in the context of the subject.

However, we believe the teaching team should take an active rather than passive attitude when a game is included. That is, we must design the course using games as the pieces of a puzzle that fit together to achieve the desired aims. A single game can lead to the teaching team adopting a passive attitude with the result that the game decides the course rather than the course deciding the game. For this reason, we believe the games must play the same role as the well-known learning aims in the e-learning environment.

On the other hand, reusing resources in the area of computer science is a standard procedure (reuse of code, programming libraries, etc.). Moreover, the concept is widely accepted and used in very different ways in the Web.

Thus, we deem it of greater interest to design the course based on reusable learning games. The nature of the game resources should be sufficiently simple, flexible, modular, and reuseable for them to act as the pieces of a puzzle.

However, not using a single source but searching for resources on the Internet is a particularly interesting added value. First, the resources must be on trustworthy sites. A large number of institutions, universities, public organizations, and solvent companies have web pages that include resources which may be very useful in this context. Thus, when using these resources students actively understand where they can and should find reliable information, a skill that will be particularly useful for their lifelong training. In addition, this method has the advantage that the teaching team has to "take the trouble" to search, analyze, and decide if the resource is the most suitable, but not only regarding content. They must focus on whether or not it is suitable for the specific profile of their students, their motivation, and so on.

In conclusion, the proposal is based on designing a course based on a set of reuseable games that meet the educational needs and fulfill the prior assumption profiles of their specific students.

Although all those needs were considered simultaneously when choosing games during the search, the choice can be reduced to three stages. The first stage of the research consisted of exploring, identifying, and evaluating the freely distributed games, paying special attention to the completeness of content. Then some games were chosen bearing in mind their instructional and motivational value. Finally, we analyzed the way they could provide added value with respect to the students' prior assumptions.

Results and Discussion — This section describes the results of the research undertaken to design the laboratory. To this end, an exhaustive web search was carried out. The next stage of our research was to explore the games available for setting up the laboratory, by classifying and choosing those that were suitable for the preliminary profile in criminology. Therefore, we conducted research into the scientific bibliography, the cybersecurity websites of both public and private institutions, installable game downloads, and lighter games.

Choosing the Games — Choosing the game laboratory components was based on the information ensuing from the research stage on the prior assumptions regarding the simulators and serious games available related to information security and their suitability for the set educational aims. The chosen games had to be accessible either by download or online, as well as facilitating active learning and individual exploration at the student's pace, and had to encourage students to search for solutions.

With these premises the laboratory work framework was based on 15 games-based activities. Each game activity was accompanied by three teaching resources: the already developed theoretical-conceptual knowledge, learning during the game, and subsequent reflection to consolidate the knowledge. This collection of games contextualized in the laboratory lets a global approach be taken to the training plan. The

	Denomination	Topics	Purposes	Provider
Game 1	Cyber CIEGE	1, 2, 3 & 6	This game can be downloaded from the CISR website. They define themselves as a simulator for teaching web security. Its modular design comprising different scenarios means that individual parts can be used as learning components. It can be used to work with the concepts of security principles, security organisation and incident management for criminology students, with sufficient technological abstraction. The activity scene unfolds in a company office and deals with the problems associated with taking on two new employees.	*P1: Center for Information Systems Security Studies and Research* (CISR) from the *Naval Postgraduate School* (NPS). Detailed information is provided on their website (cisr.nps.edu)
Game 2	The Case of the Cyber Criminal	4	It develops the concepts of vulnerabilities, threats, risks and individual crime. The learning aim of this second game is to underline the relationship between crime and cyberspace by taking advantage of the inherent predisposition of crime science students.	*P2: Federal Trade Commission* of the United States Government in collaboration with other federal agencies. Their site (onguardonline.gov)
Game 3	Auction Action	11 & 12	This is an introduction to e-commerce and its associated problems. The purpose of the activity is to draw attention to the types of fraud linked to Internet transactions	
Game 4	Follywood Squares	7 & 8	Information access-related content can be dealt with in this game; sensitive data and the reliability of the information to be found on the Internet.	
Game 5	Invest Quest	8, 11 & 12	This introduces the crime theme linked to online financial issues. The purpose of this game is to illustrate an online security scenario linked to business matters and their associated crime.	
Game 6	Mission, Laptop Security	2, 7 & 13	This strengthens the contents of physical and managed security. We need to underline the concepts related to security organisation regulations, data protection from a physical point of view and access to the information.	
Game 7	Phishing: Avoid the Bait	10, 12 and 13	This helps study the concepts of social engineering, frauds, scams, deceit, crime and how it is linked to the protection of networked data. The purpose of this activity is to focus on the security issues.	
Game 8	P2P Threeplay	3, 4 & 8	This lets us work with security areas in networks and collaboration and information sharing together with threats and their risks. It strengthens the contents of the human factor in security, malware and the threats linked to networked file-exchange scenarios.	
Game 9	Friend Finder	3, 8 & 13	This presents a scenario on the use of social networks, their threats and the necessary precautions.	
Game 10	Spam Scam Slam	10 and 12	It examines security in the use of email, hoaxes, scams and latent crime situations	
Game 11	Beware of Spyware	4 & 5	It examines the concepts of malware, its threats and incident management that is user-friendly.	
Game 12	Invasion of the Wireless Hackers	7 & 9	This is useful for strengthening the concepts of wireless networks, even introducing mobility with pervasive computing in cybersecurity.	
Game 13	ID Theft Face-off	10, 12 & 13	This introduces access to information and crime linked to identity theft online and the concepts of personal data protection.	
Game 14	Cyber Secure: Your Medical Practice	1, 2, 3 & 13	This game has more complex dynamics in a scenario that simulates a health centre with many rooms. It is directed in rounds and presents case uses with questions to be answered by the user. For our laboratory it is of interest because it deals with the concepts of security principles and organisation, the human factor and the protection of sensitive data.	*P3: Office of the National Coordinator for Health Information Technologies* (ONC), online game in several rounds: healthit.gov
Game 15	Cyberbully Zombies Attack	7 & 12	This is a simple platform game whose scene simulates a school attacked by bullies, symbolically represented as zombies. It is interesting for presenting the problems of the cyberbully and in general antisocial behaviour in cyberspace.	*P4: National Center for Missing & Exploited Children*, about awareness-raising and child protection: missingkids.com, nsteens.org

Table 3. Design model of game activities.

game activities and how they are related to course content are specified in Table 3.

Applying the Instructional Model — Having set the topics of course content and how the game activities contribute knowledge to the content, Gagné's events model and Keller's ARCS motivational model were considered in order to evaluate their characteristics and suitability for the learning process. Table 4 lists the components of the instructional model and the ARCS model that is implicit or explicit to each game.

Identifying Gagné's events in the games enabled us to set the learning conditions on which to base our instructional design and the choice of games. In addition, it allowed us to see the needs that should be provided by the teaching team to supplement the games in order to achieve the set objectives.

Taking the study, we must underline the first

Each game activity was accompanied by three teaching resources: the already developed theoretical-conceptual knowledge, learning during the game, and the subsequent reflection to consolidate the knowledge. This collection of games contextualized in the laboratory lets a global approach be taken to the training plan.

Game	Gagnés events of instruction	Keller's motivational elements
Game 1: CyberCIEGE Information Assurance & Security Policies	Events: 1, 4, 5, 6,7, 8	Attention, Relevance, Confidence, Satisfaction
Game 2: The Case of the Cyber Criminal	Events 1, 4, 6, 7, 8, 9	Attention, Relevance, Satisfaction
Game 3: Auction Action	Events: 1, 4, 6, 7, 8, 9	Attention, Relevance, Confidence, Satisfaction
Game 4: Follywood Squares	Events: 1, 6, 7, 8,9	Attention, Confidence, Satisfaction
Game 5: Invest Quest	Events: 1, 4, 7, 8, 9	Attention, Relevance, Satisfaction
Game 6: Mission Laptop Security	Events: 1, 4, 6, 7, 8, 9	Attention, Relevance, Confidence, Satisfaction
Game 7: Phishing: Avoid the Bait	Events: 1, 3, 4, 6, 7, 8, 9	Attention, Relevance, Confidence, Satisfaction
Game 8: P2P Threeplay	Events: 1, 3, 4, 6, 7, 9	Attention, Relevance, Confidence, Satisfaction
Game 9: Friend Finder	Events: 1, 3, 4, 6, 7, 9	Attention, Relevance, Confidence, Satisfaction
Game 10: Spam Scam Slam	Events: 1, 4, 6, 7, 9	Attention, Relevance, Confidence, Satisfaction
Game 11: Beware of Spyware	Events: 1, 4, 6, 7, 9	Attention, Relevance, Satisfaction
Game 12: Invasion of the Wireless Hackers	Events: 1, 4, 7, 9	Attention, Relevance, Satisfaction
Game 13: ID the Face-off	Events: 1, 4, 6, 7, 9	Attention, Relevance, Confidence, Satisfaction
Game 14: Cyber Secure Your Medical Practice	Events: 1, 4, 5, 6, 7, 9	Attention, Relevance, Confidence, Satisfaction
Game 15: Cyberbully Zombies Attack	Events: 1, 3, 4, 5, 6, 7, 9	Attention, Relevance, Confidence, Satisfaction

Table 4. Applicability values for the proposed instructional model.

event as it ensures us an introductory stimulus for "attracting the attention" of the student. In fact, the atmosphere itself outside the classroom reinforces the event: computer room, collaborative work, participation that is less formal and direct than a face-to-face master class. The evaluation carried out, however, specifically evaluates the game-element, which by playing the game shows an interest in the topic used by the game to answer and understand questions that will be answered through the act of playing it.

The second event, which deals with "reporting the aims," is not inherent to games, since, by their very nature, they are reusable pieces: serious games as an educational target. For this reason, the projection regarding the educational aims in our study plan are not informed by the game, and therefore must be supplemented by part of the teaching team through presenting the teaching aims separately and making them part of the subject teaching plan. In this respect, Gagné's number 3 event is presented, and in certain games, event 5. All of these are completed by the teaching team to strengthen the event required for learning, such as the stimulation of memory or the process guidelines.

On the other hand, specific teaching guidelines exist for each game to make up for any possible gaps. For example, should the game not offer any self-assessment, this will be included in the guidelines.

Evaluating the Laboratory — Evaluating the laboratory experience as a whole was measured using a voluntary survey. The participation index was 87.5 percent, of which 46.9 percent were male students compared to 40.6 percent female students.

The results of the choice of topics dealt with in the laboratory show a perception rated as good or excellent by 72 percent. The depth was rated as good or adequate by 92 percent, while 80 percent of the students consider the game activities to be good or excellent regarding applicability to their professional area. Also evaluated was students' perception regarding the individual

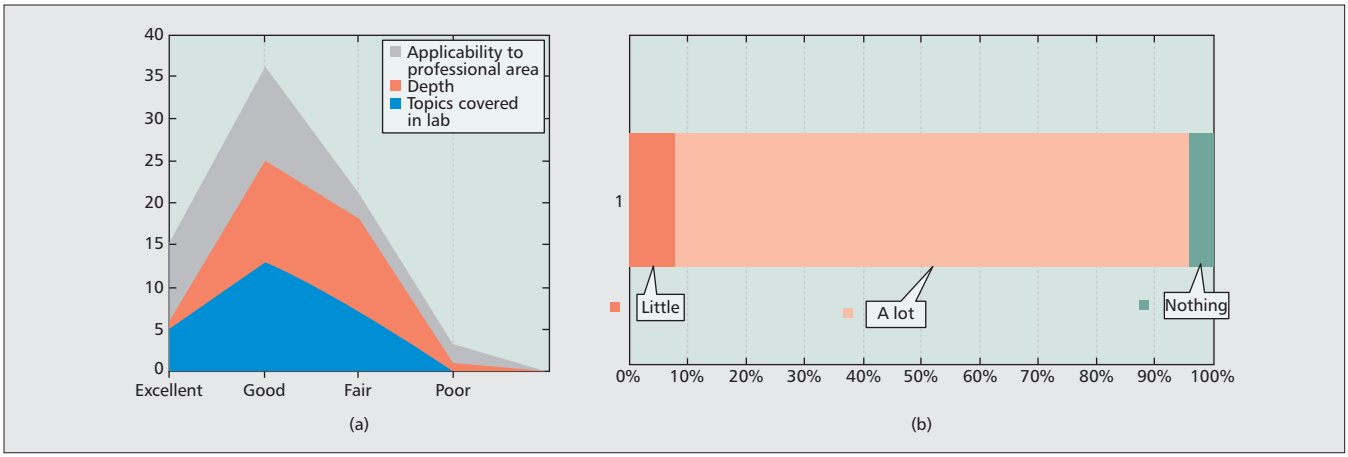


Figure 2. Representations of evaluations; a) laboratory evaluation results; b) evaluation of individual usefulness.

usefulness of the lab. A large majority consider it very useful (88 percent), while 4 percent think that it is of no use. Figure 2 shows the results in the form of a graph.

On the other hand, students were asked to put forward new topics for the laboratory. Particularly marked are their suggestions concerning issues of interest, such as social engineering, hackers, defense against and detection of attacks, insiders, cryptography in WiFi, viruses, and cyber espionage. Some of these were dealt with in some of the laboratory activities, but these suggestions reveal an interest in examining these aspects in depth after having completed the lab work on the subject.

Finally, we are of the opinion that the games laboratory enhances the learning process of the cybersecurity part of the criminology course. Students' responses were surprisingly conclusive: 96 percent responded in the affirmative.

Conclusions

In the specific case of computer-related technology, it is ever more necessary to impart knowledge to professionals who are lacking technical knowledge in this field. In this context, serious games may well turn out to be a very useful and effective tool.

However, we believe that if serious games are to be used for academic purposes, it would be highly recommendable to standardize and catalog their characteristics from both instructional and motivational points of view. If serious games are understood as learning targets, they could become a standard element of course design.

There is no doubt that engineering-based skill experiences for non-technological sciences can provide satisfactory results bearing in mind two realities: the preliminary nature of the students as digital natives and a transversal educational approach to the syllabus, which will open up bridges to engineering.

Acknowledgments

We would like to extend our gratitude to the Instituto Superior de Estudios Profesionales (ISEP CEU) and the Faculty of Law of San Pablo CEU University for their help and support.

References

[1] A. McGettrick *et al.*, "Toward Curricular Guidelines For Cybersecurity," *Proc. 45th ACM Technical Symp. Computer Science Education*, 2014.
[2] F. B. Schneider, "Cybersecurity Education in Universities," *IEEE Security & Privacy*, vol. 11, no. 4, pp. 3–4.
[3] ACM/IEEE CS, The Joint Task Force on Computing Curricula, Computer Science Curricula 2013, Ironman Draft (v. 1.0); http://ai.stanford.edu/users/sahami/CS2013/ironman-draft/cs2013-ironman-v1.0.pdf (accessed 2014).
[4] "Everyone's a Gamer — IEEE Experts Predict Gaming Will Be Integrated into More than 85 Percent of Daily Tasks by 2020," *IEEE News*, 2014; www.ieee.org/about/news/2014/25_feb_2014.html (accessed 2014).
[5] C. E. Irvine and M. F. Thompson, "Simulation of PKI-Enabled Communication for Identity Management Using CyberCIEGE'" *Proc. IEEE MILCOM*, Nov. , 2010, pp. 1758–63.
[6] H. F. O'Neil, R. Wainess, and E. L. Baker, "Classification of Learning Outcomes: Evidence from the Computer Games Literature," *Curriculum J.*, vol. 16, no.4, pp. 455–74.
[7] P. Wouters and H. van Oostendorp, "A Meta-Analytic Review of the Role of Instructional Support in Game-Based Learning," *Computers & Education*, vol. 60, no. 1, Jan. 2013, pp. 412–25.
[8] R. M. Gagné *et al.*, *Principles of Instructional Design*, 5th ed., Wadsworth.
[9] J. M. Keller, *Motivational Design for Learning and Performance: The Arcs Model Approach*, Springer.

Biographies

Jesús Cano [M] is an engineer with military rank, with experience of over 17 years in public service. At present, he works in the Constitutional Court of Spain. His research interests are about technologies for citizens. He has an M.S. in communications, networks, and content management from UNED University. He is a member of IEEE CS eGovernment Special-Technical-Community and belongs to several societies such as Communications, Computer, Education, and Signal Processing of IEEE. Contact him at jcano@scc.uned.es.

Roberto Hernández [SM] is an associate professor at the Control and Communication Systems Department at UNED. He has been Dean of the School of Computer Science at UNED for eight years. His research interests include e-government, quality of service support in distributed systems and development of infrastructures for e-learning, software quality and architecture systems engineering. He has co-authored more than 80 publications in international journals and conferences. Contact him at roberto@scc.uned.es.

Salvador Ros [SM] is a senior lecturer at UNED in the School of Computer Science. Currently, he is vice-dean of technologies at the same school. He has been director of learning technologies at UNED for six years. His research and professional activity in general is focused on enhanced learning technologies for distance learning scenarios, learning analytics, and security and privacy in communications. Contact him at sros@scc.uned.es

RADIO COMMUNICATIONS: COMPONENTS, SYSTEMS AND NETWORKS

Thomas Alexander

Amitabh Mishra

The history of radio communications has been characterized by a constant march toward ever higher frequencies. Early spark and continuous wave (CW) communications at kilometer wavelengths soon gave way to medium wave transmissions in the 500 kHz range (i.e., wavelengths of hundreds of meters). Increasing spectrum congestion and the discovery of ionospheric reflection shifted focus during the middle of the last century to the shortwave band where wavelengths are measured in meters. Satellite and terrestrial digital communications required much higher bandwidths and better spectrum sharing, pushing the use of UHF (ultra high frequency) wavelengths approaching 1 GHz. In the last two decades the explosion of cellular and high-speed digital links have caused us to reach for the centimeter band in earnest; modern cellular and wireless LAN systems use frequencies in the 2–5 GHz range (15–6 cm).

However, even the relatively wide channels (up to 160 MHz for IEEE 802.11) available at these frequencies are now proving insufficient for bandwidth-hungry mobile applications. In the past few months there has been an upsurge of interest in the millimeter-wave band, generally in the 30–60 GHz range. While these frequencies have certainly been used in the past — Jagadish Chandra Bose reportedly experimented with 60 GHz waves as far back as 1897! — it is only recently that they have been considered for applications beyond radars, sensing, and proprietary point-to-point links. The attractiveness of millimeter-wave frequencies lies, of course, in the enormous channel bandwidths they represent, for example, over 7 GHz in the 60 GHz unlicensed band alone. Also, their propagation characteristics imply a much higher degree of spectrum sharing than is possible with lower frequencies, and even large antenna arrays can fit on a postage stamp at these wavelengths.

With the high current interest in millimeter-wave communications, we felt that it would be appropriate to bring you an issue of RCS dedicated to various developments and new research in this field. We are pleased to place before you no less than three interesting articles covering work in this area.

It is well known that millimeter-wave propagation is substantially different from lower frequency bands. These characteristics are well characterized for point-to-point links, where atmospheric attenuation is perhaps the most significant issue. However, a dense urban environment represents an entirely different situation, particularly when multiple-input multiple-output (MIMO) is thrown into the mix. Our first article, "MIMO for Millimeter Wave Wireless Communications: Beamforming, Spatial Multiplexing, or Both?" deals with this new area. The article describes a ground-breaking propagation measurement campaign conducted in urban New York, analyzes some of the startling conclusions drawn from the data, and then provides a tutorial view of the insights and challenges of utilizing MIMO at millimeter-wavelengths in an urban environment.

Our second article, "MIMO Precoding and Combining Solutions for Millimeter Wave Systems," deals with efficiently utilizing the wealth of MIMO propagation modes available in millimeter-wave bands, again for typical indoor and urban areas. At the extremely short wavelengths used, it becomes economically feasible to build arrays of very large numbers of antennas; hundreds or even thousands of antennas have been proposed for overcoming antenna aperture limits. However, nothing comes for free, and driving all of these antennas or processing their signals using today's semiconductor and transceiver technologies is very expensive. The authors of the article explore a hybrid architecture that combines analog beamforming with digital processing chains and lower-resolution analog-to-digital converters, which offers the potential to utilize all those antennas quite efficiently. These techniques are likely to be of considerable interest as millimeter-wave communications becomes more ubiquitous.

The third article in this issue of our Series is an Invited Paper on the recently ratified IEEE 802.11ad standard for 60 GHz wireless LANs. Utilizing millimeter-wave frequencies involves more than understanding its propagation

characteristics and developing RF techniques to take advantage of them; we also need medium access control (MAC) and physical layer (PHY) functions that can efficiently utilize these techniques. This article includes among the authors the Chair and the Technical Editor of the IEEE 802.11ad task group, and provides an in-depth understanding of the rationale and the novel techniques written into the IEEE 802.11ad standard to harness the benefits of millimeter-wave communications while at the same time coping with its many issues. As IEEE 802.11ad represents the first standardized millimeter-wave radio access network technology, its progress and evolution in the industry will undoubtedly be watched closely.

We hope you will find our issue on millimeter-wave systems of interest and benefit, and look forward to providing more high-quality tutorials and research summaries in future issues. As always, of course, we solicit your suggestions and feedback, and also would like to invite you to continue to submit tutorial papers dealing with recent research and developments to the Radio Communications Series. As channel bandwidths creep ever higher, perhaps some day in the future we will be able to bring you articles on terahertz communication systems!

BIOGRAPHIES

THOMAS ALEXANDER [M] (talexander@ixiacom.com) is a senior architect at Ixia. Previously, he worked at VeriWave Inc. (acquired by Ixia), PMC-Sierra Inc., and Bit Incorporated (acquired by PMC-Sierra), and prior to that was a research assistant professor at the University of Washington. He has been involved in various aspects of wired and wireless networking R&D since 1992, in the areas of ATM, SONET/SDH, Ethernet, and (since 2002) wireless LANs. He is also active in standards development, and has served as an Editor of IEEE 802.3ae, Chief Editor of IEEE 802.17, and a Technical Editor of IEEE 802.11. He received his Ph.D. degree from the University of Washington in 1990.

AMITABH MISHRA [SM] (amitabh@cs.jhu.edu) is a faculty member with the Information Security Institute of Johns Hopkins University, Baltimore, Maryland. His current research is in the area of cloud computing, data analytics, dynamic spectrum management, and data network security. In the past he has worked on the cross-layer design optimization of sensor networking protocols, media access control algorithms for cellular–ad hoc interworking, systems for critical infrastructure protection, and intrusion detection in mobile ad hoc networks. His research has been sponsored by NSA, DARPA, NSF, NASA, Raytheon, BAE, APL, and the U.S. Army. In the past, he was an associate professor of computer engineering at Virginia Tech, and a member of technical staff with Bell Laboratories working on the architecture and performance of communication applications running on the 5ESS switch. He received his Ph.D. in electrical engineering from McGill University. He is a member of ACM and SIAM. He has written 80 papers that have appeared in various journals and conference proceedings, and holds five patents. He is author of a book, *Security and Quality of Service in Wireless Ad hoc Networks* (Cambridge University Press, 2007) and a Series Editor of *IEEE Communications Magazine*.

MIMO for Millimeter-Wave Wireless Communications: Beamforming, Spatial Multiplexing, or Both?

Shu Sun, Theodore S. Rappaport, Robert W. Heath, Jr., Andrew Nix, and Sundeep Rangan

ABSTRACT

The use of mmWave frequencies for wireless communications offers channel bandwidths far greater than previously available, while enabling dozens or even hundreds of antenna elements to be used at the user equipment, base stations, and access points. To date, MIMO techniques, such as spatial multiplexing, beamforming, and diversity, have been widely deployed in lower-frequency systems such as IEEE 802.11n/ac (wireless local area networks) and 3GPP LTE 4G cellphone standards. Given the tiny wavelengths associated with mmWave, coupled with differences in the propagation and antennas used, it is unclear how well spatial multiplexing with multiple streams will be suited to future mmWave mobile communications. This tutorial explores the fundamental issues involved in selecting the best communications approaches for mmWave frequencies, and provides insights, challenges, and appropriate uses of each MIMO technique based on early knowledge of the mmWave propagation environment.

INTRODUCTION

Multiple-input multiple-output (MIMO) techniques have been deployed in modern wireless networks to improve reliability and capacity. The multiple antennas in a MIMO system may be leveraged in different ways. Among the most common approaches are spatial multiplexing (SM) and beamforming (BF). These technologies are already in use, or are planned, for wireless local area network (WLAN) and 4G Long Term Evolution (LTE) cell phone systems. They are certain to play a significant role in the 5th generation (5G) mobile systems that are envisaged to be deployed in the coming years [1–4]. In this tutorial, we compare and contrast the capacity potential for transmit SM and BF approaches, and explore how they may be best used in future millimeter-wave (mmWave) wireless systems with channel bandwidths that will exceed hundreds of megahertz.

SPATIAL MULTIPLEXING

SM provides multiplexing gain that increases communication system throughput by subdividing an outgoing signal stream into multiple pieces, where each piece is transmitted simultaneously and in parallel on the same RF channel through different antennas. The slightly different propagation paths cause each of those pieces — called spatial streams — to "see" a different frequency-selective time-dispersive channel. Using the observations from multiple antennas at an individual receiver (single-user MIMO, SU-MIMO) or when multiple end user receivers are simultaneously receiving signals from a MIMO transmitter (multi-user MIMO, MU-MIMO), the receiver can reconstruct the original transmitted sequence by using an estimate of the channel state information (CSI). If the receiver's CSI is also available at the transmitter, precoding can be employed at the transmitter to send the data streams along the best dimensions of the channel, resulting in higher throughput with low-complexity receivers. Using channel knowledge at the transmitter is called *closed-loop* SM, where feedback is usually used to convey CSI from the receiver back to the transmitter [5]. In time-division duplex (TDD) single-frequency systems, the transmitter can directly measure the CSI when the receiver decides to transmit without the need for feedback, allowing *open-loop* SM. Arogyaswami Paulraj and Thomas Kailath proposed the concept of SM using MIMO in 1993 (U.S. Patent 5,345,599). Today, IEEE 802.11n access points (APs) must be capable of implementing at least two spatial streams using SM, up to a maximum of four, and IEEE 802.11ac extends those air interface concepts further with wider RF channel bandwidths (up to 160 MHz), and up to eight MIMO spatial streams and as many as four downlink MU-MIMO clients with high-order modulation (up to 256-quadrature amplitude modulation, QAM). Both open-loop and closed-loop SM are supported in wide area network standards such as IEEE 802.16e/m, 4G cellular LTE, and LTE-Advanced (LTE-A) [1, 6]. For SM to work well, the channel must provide sufficient decorrelation between the different closely spaced antennas, but even if the channel does not have sufficient diversity to support SU-MIMO, MU-MIMO can still provide increased capacity on a single radio channel to multiple users dispersed in space.

Shu Sun, Theodore S. Rappaport, and Sundeep Rangan are with NYU WIRELESS and New York University.

Robert W. Heath, Jr. is with WNCG and the University of Texas at Austin.

Andrew Nix is with the University of Bristol.

0163-6804/14/$25.00 © 2014 IEEE

Beamforming

BF is a classical array signal processing technique where multiple antenna elements are adaptively phased to form a concentrated and directed beam pattern [7, 8]. BF can be used at both the transmitter (TxBF) and receiver (RxBF) to provide significant array gains, thereby providing increased signal-to-noise ratio (SNR) and additional radio link margin that mitigates propagation path loss. Also, BF provides reduced co-channel interference from the spatial selectivity of the directional antennas [7, 9, 10]. Multiple beams may be combined at the receiver, decreasing the path loss even further [9]. For mmWave systems, BF offers great promise since highly directional adaptive antennas can be made in very small form factors, and steered in various directions to exploit reflections and scattering from objects for maximal signal strength while coherently aligning the received waveforms [1, 9, 11]. While the zero-forcing beamforming (ZF-BF) used in closed-loop SM is also considered in the literature to be a type of BF, that method simply seeks to form antenna beams which dampen signals in undesired directions by achieving orthogonality in the channel matrix (H matrix) without optimizing the radiated antenna energy in a particular desired direction. Here we use the term *beamforming* more classically to refer to a directed steered beam, as described in [7, 8]. Several BF architectures are reported in the literature, as shown in Fig 1. Analog BF includes a single RF chain with $N_x \times N_z$ analog phase weights (one per antenna element) to focus the BF gain in the direction of the dominant channel path or paths. Digital BF requires $N_x \times N_z$ RF chains (one per antenna element), where the per-element BF weights are applied digitally to provide the best matching to the dispersive channel for a given antenna size.

Combination of Various MIMO Techniques

Key factors in determining when to use SM or BF are the operating SNR and the equivalent bandwidth provided to each user. SM uses multiple streams on a single carrier to increase the capacity per user, but is most effective when radio links operate in a high SNR regime and are bandwidth-limited (and not power-limited). Furthermore, SM is only effective when the channel provides sufficient diversity, or rank (effectively, the number of streams that can be supported by the MIMO H matrix). At low SNR operating points (e.g., power-limited channels with little interference), SM provides little benefit since the transmitter must split its power across the different spatial streams, thus weakening each stream and inducing bit errors that limit overall capacity gains. Thus, in the power-limited regime, BF may provide greater capacity by increasing SNR to allow the use of higher order modulations. Early work shows that mmWave systems may operate in either regime, but will often be power-limited rather than bandwidth-limited (due to much greater spectrum allocations and higher path loss associated with mmWave wavelengths), and will also often be noise-limited rather than interference-limited due to the use of BF to avoid co-channel interference while exploiting angle diversity [1, 2, 7, 9, 12, 13]. As mmWave communication systems evolve, they will exploit much wider RF channels having 500 MHz bandwidth or more, and devices will use dozens of antennas due to the much smaller wavelengths involved, thereby offering many dimensions in the channel that can be leveraged for both SM and BF purposes [1, 4, 11, 14]. The benefits of spatial multiplexing and beamforming can be achieved at the same time with the proper hardware architecture. For example, it is possible to use multiple beams in BF to increase SNR in power-limited situations, while also providing unique data streams on each of the beams on the same carrier frequency to increase user data rates, as long as the mmWave channel has enough sufficiently differ-

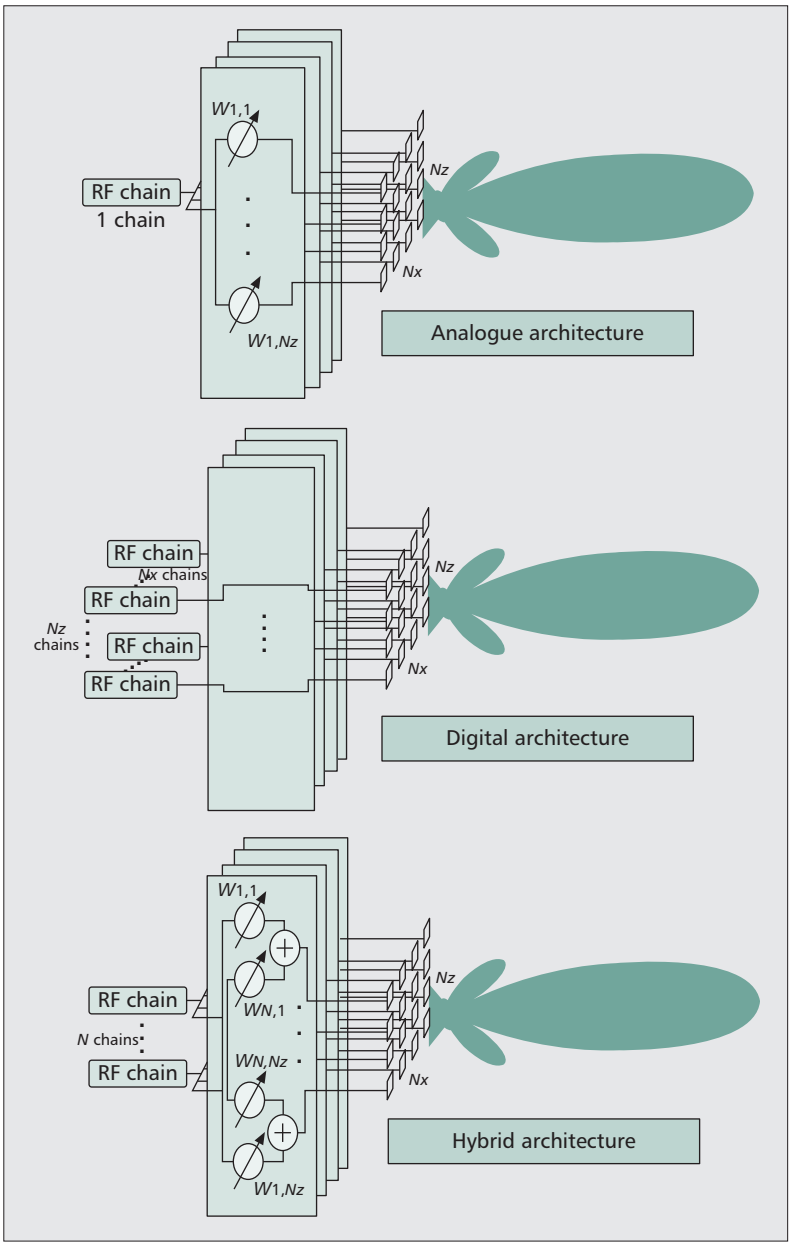


Figure 1. Possible RF architectures for mmWave systems, each featuring different hardware, power, and complexity trade-offs.

To understand how SM and BF will work at mmWave frequencies, propagation measurements are required, yet to date, there has been little work aimed at understanding mmWave wireless communication signals and propagation characteristics that incorporate MIMO.

ent propagation paths in the spatial and polarization domains. This ability to simultaneously exploit BF and the multiple streams of SM is a wonderful situation not previously available to UHF/microwave wireless networks that currently use low-gain or omnidirectional antennas.

Furthermore, if the channel can provide multiple spatial degrees of freedom, where each unique beam has both strong propagation path and small root mean square (RMS) delay spread (or just a few significant multipath components [MPCs]), it becomes possible to use both SM and BF with a simplified receiver architecture using simple time domain equalization or rake receivers (e.g., wideband single carrier modulation techniques) over very wideband channels at greatly reduced latency. This is in contrast to today's cellular and WLAN modulations, which use multi-carrier frequency domain equalization with small subcarrier spacing to create narrowband flat-fading channels for MIMO exploitation. We now explore the joint use of SM and BF in mmWave channels to provide insight into system performance and challenges.

Figure 1 shows three different BF architectures attached to a 2D antenna array. Analog BF includes a single RF chain with $N_x \times N_z$ analog phase weights (one per antenna element). The digital BF requires $N_x \times N_z$ RF chains (one per antenna element). Here the per-element BF weights are applied digitally. Digital BF offers the greatest flexibility, but is also the most demanding in terms of power and cost — a particularly important concern for mmWave front-ends due to the very wide bandwidths and large numbers of antenna elements normally involved. Implementing both SM and BF can be achieved with a hybrid beamforming architecture. This method uses N RF chains ($N \ll N_x \times N_z$), each connected to $N_x \times N_z$ analog phase weights. Hybrid BF reduces the number of required RF chains and enables N multiple data streams to be sent in different spatial directions.

Outdoor mmWave Channel Measurements and Channel Modeling Using Directional Beams

Recent work at the University of Bristol shows that at UHF/microwave frequencies, small cell mobile users are most likely to exploit just two spatial streams using SU-MIMO SM, even when the base station supports up to 16 antenna elements, due to relatively low multipath angle spreads and prevalent line of sight (LOS) channels. However, MU-MIMO at UHF/microwave frequencies achieves much better capacity than SU-MIMO, as randomly dispersed users and relatively low-gain antennas provide a channel with higher effective rank. To understand how SM and BF will work at mmWave frequencies, propagation measurements are required, yet to date, there has been little work aimed at understanding mmWave wireless communication signals and propagation characteristics that incorporate MIMO, with the earliest studies reported in [1, 3, 16].

The NYU WIRELESS research center has been studying wideband 28 and 73 GHz outdoor channels in both LOS and non-LOS (NLOS) environments in downtown New York City using an exhaustive search for possible propagation paths with steerable directional antennas as a way to study MIMO viability [3, 9, 13, 15, 16]. A broadband sliding correlator channel sounder was used to probe the propagation channel with 2.5 ns time resolution by using a length 2047 binary phase shift keying (BPSK) pseudorandom noise (PN) sequence clocked at 400 megachips per second Mc/s), thereby producing an 800 MHz first null-to-null RF bandwidth spread spectrum signal centered around the carrier frequency. The 28 GHz measurements were performed using 24.5 dBi (10.9° half-power beamwidth (HPBW)) rotatable horn antennas at both the transmitter (TX) and receiver (RX). Three TX sites were selected (two are 7 m high, and the other is 17 m high) in the 28 GHz measurements. All three TX sites used the same set of 25 RX sites. Each RX site placed an antenna 1.5 m above ground level. The 73 GHz measurements were conducted at five TX locations and 27 RX locations with a total of 74 TX-RX combinations; four TX locations were 7 m in height and the other 17 m high.

Two types of 73 GHz measurements were carried out at almost all the RX locations:

- Base-station-to-mobile scenario where the RX height was 2 m
- Backhaul-to-backhaul scenario where the RX height was 4.06 m

In the 73 GHz measurements, rotatable directional horn antennas with 7° HPBWs and 27 dBi gains were used at both the TX and RX. In both measurement campaigns, at each location, the horn antennas were systematically rotated over complete azimuth sweeps at both the receiver and transmitter, and sweeps were repeated at various elevation angles to characterize the 3D channels. Maximum measurable path loss values at 28 and 73 GHz were 178 and 181 dB, respectively. Further details on the measurement equipment, procedures, and resulting measurements and models can be found in [3, 12, 13, 14, 15, 16].

From the 28 GHz NYU WIRELESS data, the maximum RMS delay spread values in LOS and NLOS locations are 309.6 and 454.6 ns, respectively, and the maximum 20 dB down maximum excess delay (MED) times relative to the first arriving signal energy in LOS and NLOS locations are 1291.4 and 1387.4 ns, respectively. For the 73 GHz urban outdoor channel, the maximum RMS delay spread values in LOS and NLOS locations are 219.2 and 248.9 ns, respectively, and the max 20 dB down MED in LOS and NLOS locations are 762.3 and 1053.0 ns, respectively. Ray tracing shows these late arriving signals came from large surface scattering and reflections from large buildings, even though there was no line of sight path. Tables 1 and 2 summarize the TX-RX separation distance, RMS delay spread, maximum 10 dB down excess delay times, and path loss among all antenna pointing angles measured for each NLOS location in the 28 and 73 GHz outdoor measurements, respectively. The values of the maximum

TX ID	RX ID	TX-RX distance (m)	RMS delay spread (ns)			Maximum 10 dB down excess delay (ns)			Path loss (dB)			Minimum simultaneously observed RMS delay spread and path loss		Number of beams with low PL and few MPC	
			Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	RMS delay spread (ns)	Path loss (dB)	Number	Percentage of all viable measured angles (%)
COL 1	RX13	133	1.1	16.4	82.4	1.4	48.9	251.2	138.0	150.0	171.0	2.1	151.5	0	0.0
	RX14	162	1.4	43.8	107.6	3.1	114.7	274.1	142.2	150.7	166.3	1.4	166.3	0	0.0
	RX17	82	1.6	5.5	18.5	4.9	23.9	98.8	151.5	157.3	167.3	1.6	164.5	1	10.0
	RX2	61	1.1	17.2	454.6	1.0	50.5	1246.3	131.1	145.0	171.4	1.5	155.4	0	0.0
	RX4	118	1.0	20.3	171.8	1.5	63.8	409.8	146.1	153.2	168.7	1.2	155.3	5	10.6
	RX5	114	1.1	16.2	247.4	1.0	41.9	347.6	126.2	139.8	174.2	1.6	128.2	22	7.8
COL 2	RX13	138	1.1	7.6	95.5	1.4	27.7	479.4	131.3	144.7	169.4	1.6	132.2	3	2.6
	RX14	169	1.5	6.5	37.6	1.6	24.6	110.4	153.8	158.5	172.0	2.5	155.4	7	25.9
	RX17	112	1.3	17.4	43.6	2.0	60.4	148.1	148.0	154.4	170.1	1.3	165.6	0	0.0
	RX2	74	1.1	11.3	44.5	1.0	35.4	144.9	130.2	144.0	173.2	1.6	149.6	1	0.6
	RX3	143	0.9	17.1	231.2	1.3	52.4	717.3	127.9	142.8	171.0	0.9	129.7	7	2.7
	RX4	156	1.1	16.2	37.2	1.6	51.1	128.6	149.3	154.1	169.1	1.2	153.6	4	19.0
	RX5	150	0.9	10.5	83.0	1.9	32.3	185.5	132.8	145.5	167.9	1.8	132.8	1	0.8
KAU	RX10	77	0.9	29.5	244.9	1.2	93.3	592.3	132.7	145.8	172.4	1.0	159.1	3	2.0
	RX12	120	1.2	22.5	58.4	3.1	73.7	168.2	150.2	157.2	167.6	1.2	163.6	5	9.3
	RX14	82	1.0	15.1	97.4	1.1	59.0	298.6	138.1	146.6	170.9	1.0	156.6	0	0.0
	RX16	97	1.3	2.1	3.9	1.7	8.3	16.5	147.7	154.7	171.7	2.0	147.7	10	35.7
	RX18	187	1.2	8.9	158.0	0.9	22.1	365.6	156.7	162.6	173.7	1.2	163.6	15	62.5
	RX19	175	1.0	17.6	296.5	1.4	40.7	823.4	147.3	156.5	173.2	1.0	153.1	9	10.6
	RX21	117	0.9	28.3	198.5	1.4	71.8	502.1	128.1	141.2	173.5	1.0	128.1	14	7.2

Table 1. TX-RX separation distance, RMS delay spread, maximum 10 dB down excess delay, and path loss among all antenna pointing angles for each TX-RX location combination in the 28 GHz outdoor NLOS measurements in New York City. These measurements were conducted in 2012 using three discrete elevation pointing angles [3], rather than finding the strongest elevation angles as was done for the 73 GHz measurements in 2013 [16] given in Table 2. The minimum combined RMS delay spread and path loss is obtained by selecting the smallest value of [10 × RMS delay spread (ns) + path loss (dB)] over all measured TX-RX pointing angles at each location. "Low PL" and "few MPC" denote path loss within 10 dB of the lowest measured value at each location and less than or equal to three multipath components, respectively. All the values are extracted from all pointing angles at both the TX and RX that had recoverable signals.

values of 20 dB down MED at each location are usually not much greater than the maximum 10 dB down excess delay values, and are thus not shown.

Tables 1 and 2 also illustrate the results of a simple algorithm to find beam pointing directions that can simultaneously minimize both RMS delay spread and propagation loss (e.g., finding the best single beam pointing directions at the TX and RX for both maximum SNR and minimum MPCs for minimal equalization). The minimum combined RMS delay spread and path loss values in the tables are obtained by searching over all measured angles at each location and selecting the particular TX-RX pointing angle combination that yields the smallest value of [10 ×RMS delay spread (ns) + path loss (dB)]. While this equation is quite simple, it offers an example of a fast

TX ID	RX ID	TX-RX distance (m)	RMS delay spread (ns)			Maximum 10 dB down excess delay (ns)			Path loss (dB)			Miminum simultaneously observed RMS delay spread and path loss		Number of beams with low PL and few MPC	
			Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	RMS delay spread (ns)	Path loss (dB)	Number	Percentage of all viable measured angles (%)
	RX1	48	0.8	22.7	202.8	2.0	60.5	485.9	136.6	146.3	176.5	0.9	138.2	30	16.0
	RX10	95	0.9	7.1	84.7	1.9	24.5	252.3	141.9	155.2	176.1	1.0	142.5	8	8.3
	RX11	109	1.2	6.5	26.3	1.8	19.9	92.6	156.0	164.9	176.5	1.2	164.2	5	20.0
	RX12	140	0.8	10.7	42.7	1.3	30.6	147.3	145.7	153.8	171.4	1.0	149.7	9	19.6
	RX2	53	0.9	12.1	200.3	1.2	40.3	1053.0	142.5	155.4	177.3	1.4	145.9	6	6.4
COL1	RX3	64	0.8	11.3	113.8	1.1	38.0	522.4	137.7	150.0	179.1	0.8	138.4	21	11.2
	RX4	104	1.0	5.2	113.3	1.3	17.1	347.7	139.0	151.9	177.9	1.2	142.7	7	5.3
	RX5	95	1.0	9.0	212.9	0.9	26.3	817.1	136.7	150.5	179.6	1.4	137.3	9	7.5
	RX6	71	1.2	16.9	248.9	1.1	45.6	628.9	142.0	155.2	177.6	2.0	142.0	3	3.4
	RX7	76	1.1	7.6	79.8	1.0	23.3	182.8	142.5	156.2	178.2	2.3	142.5	6	6.3
	RX9	58	1.0	5.6	53.2	1.4	15.5	146.8	143.4	155.5	178.0	2.2	145.9	2	2.3
	RX1	88	0.9	6.8	31.6	1.3	24.5	130.6	137.8	149.9	175.7	2.0	138.2	2	3.1
	RX10	137	1.0	9.0	156.6	1.3	29.6	630.7	144.1	154.8	175.2	1.5	144.1	7	7.6
	RX11	148	1.8	1.9	2.3	2.6	5.4	8.1	166.3	168.1	173.0	1.9	167.0	7	100.0
	RX12	145	1.1	9.1	205.9	2.1	32.4	840.2	148.1	157.8	173.8	1.1	148.7	7	11.1
	RX2	91	1.1	13.3	145.2	1.3	44.2	422.8	151.6	158.7	175.6	1.1	159.5	4	3.5
COL2	RX3	106	0.9	8.1	173.4	0.9	27.7	477.9	143.7	154.4	176.8	1.3	145.9	6	5.1
	RX4	139	1.1	8.3	38.7	1.4	27.8	140.6	154.6	163.5	176.5	1.1	161.7	3	6.5
	RX5	128	0.9	10.3	83.7	1.2	30.9	198.3	154.6	162.3	177.1	0.9	154.6	4	6.8
	RX6	99	1.4	7.5	79.7	1.1	28.0	391.9	153.6	160.5	176.1	2.2	157.2	8	17.8
	RX7	107	1.5	9.5	21.0	1.3	33.7	77.9	158.3	162.4	175.4	1.5	166.9	8	33.3
	RX9	70	1.1	6.9	31.0	1.6	26.2	133.4	150.6	158.4	175.7	1.1	162.7	2	4.3
	RX15	80	0.9	15.8	166.4	0.9	46.5	425.3	144.2	152.3	173.7	0.9	157.1	1	1.2
	RX18	59	0.8	13.2	97.7	0.9	38.6	390.2	125.9	143.0	180.5	1.8	129.1	12	2.4
	RX19	79	1.1	13.7	158.2	0.9	44.9	432.6	137.2	152.7	182.0	1.1	156.5	1	0.3
	RX20	168	1.0	7.0	157.7	1.4	19.9	360.8	139.3	155.4	175.4	1.0	139.3	4	4.3
KAU	RX21	181	0.8	4.2	83.3	1.1	12.4	180.1	141.7	152.7	172.1	1.0	148.5	22	13.2
	RX22	129	0.9	15.2	182.6	1.4	43.2	375.8	134.6	148.7	171.9	1.4	134.6	9	6.3
	RX23	127	0.9	7.3	86.1	0.9	23.0	248.0	142.0	154.1	172.3	1.2	142.0	6	5.6
	RX24	118	0.8	6.8	76.5	0.9	20.3	185.2	121.2	141.7	174.2	1.0	121.2	3	1.3
	RX25	117	0.9	9.5	74.2	0.9	28.7	190.9	142.1	154.8	178.0	0.9	157.2	3	1.4
	RX25	190	1.0	2.1	5.8	0.9	5.2	20.8	153.3	164.3	178.5	1.1	153.3	21	18.9
KIM1	RX26	50	0.9	8.3	107.4	1.2	21.6	407.6	137.4	151.5	175.1	0.9	137.4	13	6.6
	RX27	74	1.1	9.2	124.2	0.9	36.6	983.6	146.1	154.7	175.5	1.1	163.5	0	0.0
KIM2	RX25	182	1.2	4.6	79.8	1.8	11.9	229.8	160.2	167.1	175.5	1.2	160.2	20	46.5

Table 2. TX-RX separation distance, RMS delay spread, maximum 10 dB down excess delay, and path loss among all antenna pointing angles for each TX-RX location combination in the 73 GHz outdoor NLOS measurements in New York City [16]. The minimum combined RMS delay spread and path loss is obtained by selecting the smallest value of [10×RMS delay spread (ns) + path loss (dB)] over all measured TX-RX pointing angles at each location. "Low PL" and "few MPC" denote path loss within 10 dB of the lowest measured value at each location and less than or equal to three multipath components, respectively. All the values are extracted from all pointing angles at both the TX and RX that had recoverable signals.

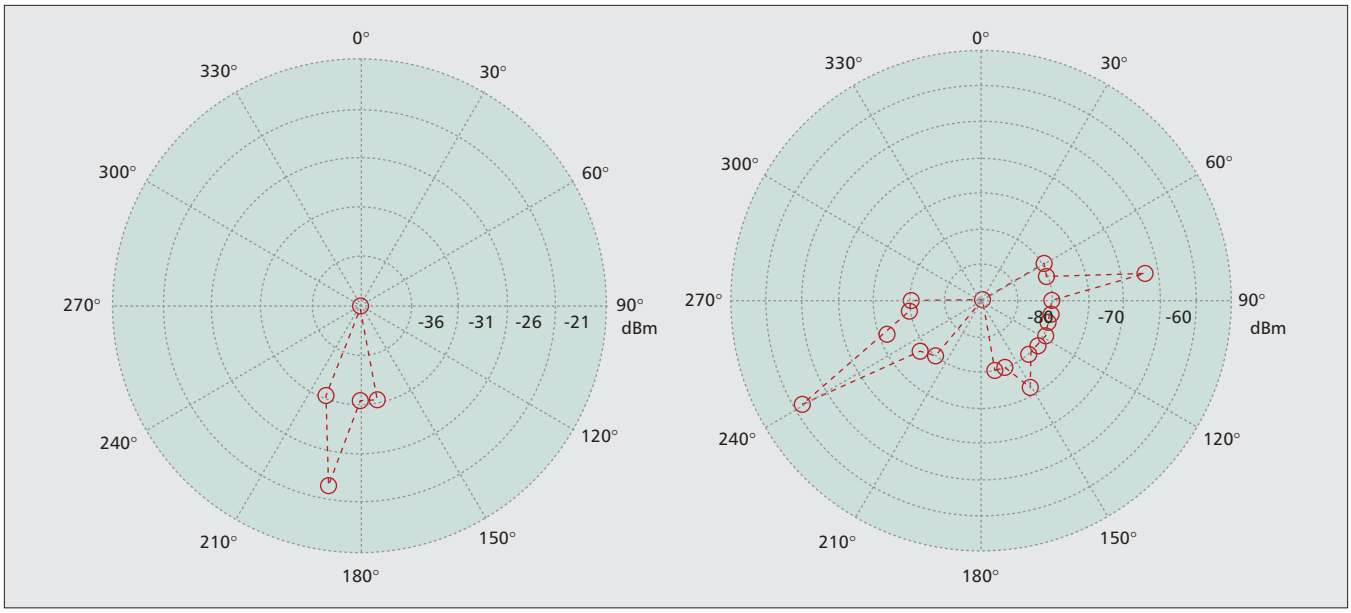


Figure 2. Example of spatial azimuth angle of departure (AoD) polar plots in LOS (left) and NLOS (right) environments at 28 GHz, as derived from the outdoor statistical channel model developed by NYU WIRELESS [14].

method of determining pointing angles with both low RMS delay spread and low path loss (where beamforming with little/no equalization could be used). More intelligent algorithms can be used to rapidly process a beam searching procedure to determine the "best" pointing angles at both the transmitter and receiver that yield both minimum time dispersion and path loss. The rightmost two columns of the tables show the number of unique beam pointing combinations at the TX and RX that simultaneously provided low path loss (no more than 10 dB larger path loss than the corresponding minimum path loss measured at the location with the strongest pointing angles at the TX and RX) and with three or fewer resolvable MPCs. These values were extracted from all of the measured pointing angles at the TX and RX that had recoverable signals. The percentages shown in the tables are obtained at each location by dividing the number of beam pointing angle pairs that offer both low path loss (within 10 dB of the absolute minimum at each location) and three or fewer MPCs by the total number of unique beams that were measured over all angles where received power was detected, thus showing the percentage of "golden" pointing directions that offer low-loss low-MPC beams among all of the measured beams.

A small RMS delay spread indicates that the received power is contained only in a few multipath components arriving closely in time. By selecting a beam with both low RMS delay spread and path loss, relatively high power can be received without complicated signal processing (e.g., equalization) at the receiver. It can be seen that beams with both low path loss and a small number of MPCs exist at most of the TX-RX location pairs, indicating that wideband signals could exploit narrow beamwidths and single-carrier modulations that do not suffer peak-to-average-power ratio (PAPR) penalties, while using no equalization whatsoever, or simple equalizer or rake receiver structures.

Works in [1, 9, 14, 15] show that receivers in NLOS locations in an urban core can expect, on average, between two to four distinct spatial beams with strong received powers — these beams can be combined to increase the total received power. Similarly, two to three spatial beams may be launched at the transmitter in NLOS conditions, although LOS channels generally offer only one dominant transmit beam (Fig. 2). Beam combining, which refers to combining the received powers recorded from different pointing angles measured at a particular TX-RX location pair, was theoretically studied at both 28 and 73 GHz based on the measured data [9]. Coherent equal-gain combining is considered here, which means that received powers from each of the different beam headings of interest are combined using known carrier phase (and time delay) information (resulting in optimal/maximum power from each of the combined beams). The results show that beam combining can significantly reduce the propagation path loss exponent (PLE), for example, using a 1 m free space reference distance, from 4.6 (46 dB per decade of distance loss) to 3.2 (32 dB per decade of distance loss), and also reduces the log-normal shadow fading from 11.3 to 7.2 dB, thus improving SNR and extending link coverage. By coherently combining the four strongest signals from four distinct beams, around 28 dB of link improvement is achieved when compared to an arbitrarily pointed single beam over a 100 m TX-RX separation at 73 GHz, and more than 10 dB of improvement when compared to a single optimum beam. The maximum possible improvement on received power at 28 GHz reaches 24 dB over an arbitrarily pointed beam based on propagation measurements. Thus, the use of BF with smart antennas to exploit the spatial degrees of freedom in the mmWave propagation channel is quite promising for outdoor urban NLOS channels [8, 9].

A 3D Third Generation Partnership Project

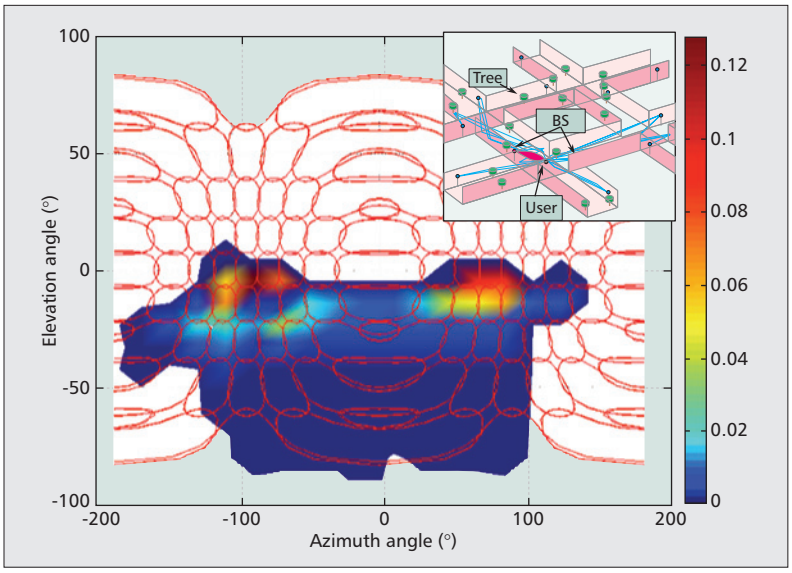


Figure 3. PDF of the spatial spread of the dominant multipath component as seen from the base station in our outdoor street scenario. Overlaid are the 3 dB contours of the beamforming codebook.

(3GPP)-style statistical spatial channel model (SSCM) has been built from the extensive 28 and 73 GHz New York City propagation database, where ray-tracing (RT) techniques were used to align the exact propagation times at different pointing angles. The model faithfully recreates the measured data, and provides statistical models for omnidirectional receiver antennas (so that arbitrary directional antennas may be applied easily), as well as path loss models, time cluster statistics, and spatial lobes that contain multiple time clusters (an observed situation not represented in past cellular standards) [14, 15]. The SSCM supports realistic outdoor MIMO performance investigations and can be used in system capacity evaluation, such as conducted in [13] for microcellular deployments with three-way sectored cells and a 200 m inter-site distance (100 m cell radius). The New York City measurements and resulting simulations show that arbitrary receiver locations often have at least two to four significant spatial pointing directions [1, 14, 15]. As found in [13], even at an RF channel bandwidth of 1 GHz, a surprising majority of the links were bandwidth-limited, operating at sufficiently high SNR to benefit from SM as well as from BF. If the inter-site distance were increased, we would expect SM gains to diminish, although BF gains would still be available.

The lack of significant diffraction at mmWave frequencies allows researchers and designers to rely on ray-tracing to determine suitable MIMO approaches [1]. At the University of Bristol, a 3D outdoor mmWave ray-tracing tool has been developed based on a Manhattan-style city grid (Fig. 3). The model has been used to analyze the 3D angle distribution from the base station (BS) to the mobile clients. Figure 3 shows the probability density function (PDF) of the elevation and azimuth angles of the strongest multipath component departing the BS to each of the users. As described below, BF is commonly implemented using codebooks, where each entry represents a specific beam pattern. By way of example, if we take the IEEE 802.15.3c codebook standard [1, Ch. 9] for an 8 × 8 BS array, we can plot the contours where the power of the beam associated with each codebook entry drops by 3 dB relative to its peak (red lines). This allows us to "see" how the codebook beams overlay on the azimuth and elevation angle distributions.

The PDF of the users' azimuth angle is not uniformly distributed; instead, it is clustered within a confined azimuth spread centered about +/–90°. This trend is further emphasized if users are constrained to the sidewalks. Angular statistics are important since, in the above scenario, the BS array would select the *same* beam pattern for *many* users. Users located in the same exclusive region (as defined by a specific beam contour) must be separated in polarization, time, and/or frequency to avoid interference. Polarization is one promising approach to ensure MIMO diversity at mmWave frequencies, even in LOS channels [12].

Outdoor mmWave MIMO Systems

While the performance of MIMO systems has been intensively investigated for radio frequencies below 6 GHz where relatively few low-gain antennas are used at the mobile device, their characteristics have barely been studied at mmWave frequencies. In this section, we consider mmWave MIMO from a system's perspective, explore how BF will work using feedback, and then examine the performance of MIMO communication in the mmWave channel by studying mutual information distributions. Finally, we consider the benefits of antenna polarization to increase rank.

Feedback and Beamforming Approaches

CSI at the transmitter, when used appropriately, achieves the highest performance of any MIMO system. The two classic approaches for obtaining CSI at the transmitter are through channel reciprocity and feedback.

Reciprocity is exploited in TDD systems where the same frequency is used for transmission alternately in both directions of the communication link. Because the propagation paths are reciprocal, the channel seen between a transmission by node A received at node B is the transpose of the channel seen between a transmission by node B received at node A. Exploiting channel reciprocity requires additional calibration mechanisms to account for asymmetry in the RF paths. TDD and reciprocity are used in cellular systems, for example, the TDD mode of 3GPP LTE, or as contemplated at 73 GHz in [12].

Feedback can be used more generally in any communication system where there is two-way communication. In such systems, the receiver sends its estimate of the CSI back to the transmitter over the feedback link. Because feedback bandwidth is limited, the CSI is often further

$$W(n_x, n_z, k_x, k_z) = j^{\text{fix}\left\{\frac{n_x \times \text{mod}[k_x + (K_x/2), K_x]}{K_x/4}\right\} + \text{fix}\left\{\frac{n_z \times \text{mod}[k_z + (K_z/2), K_z]}{K_z/4}\right\}} \quad (1)$$

compressed so that only the essential information is sent. IEEE 802.11n supports both uncompressed and compressed forms of feedback. 3GPP LTE supports what is known as *limited feedback*, where subspace information about the channel is quantized using a codebook that is known to both the receiver and the transmitter [5].

Both reciprocity and feedback pose challenges in mmWave systems due to the large numbers of antennas and RF chains involved. For a system attempting to exploit reciprocity, calibration may require too many additional circuits or extensive system overhead to implement efficiently. If the system exploits feedback, the overhead may be unacceptably high as the amount of feedback required grows with the number of antennas. Beam training is an alternative approach that leverages feedback, which is more tailored toward exploiting features such as reduced dimensionality or sparsity found in mmWave MIMO channels. Beam training is used to configure transmit and receive beamforming vectors in IEEE 802.11ad. Essentially, the transmitter sends information on several alternate beams and uses feedback from the receiver to determine the best beam. Subsequent iterations are performed over progressively narrower sets of beams [1, Ch. 8]. Beam training is well suited for BF, especially when the BF is performed in analog, and an open area of research is to develop beam training protocols or other feedback techniques for closed-loop SM with multiple streams and multiple beams. For illustration purposes, here we describe three attractive beamforming approaches in more detail: codebook-based BF, BF weight optimization based on angle of departure (AoD) estimation, and long-term BF.

Codebook-Based Beamforming — A set of phase weights (one per antenna element) can be combined to form a specific codebook entry. A number of these entries can be grouped to form a codebook. Different codebooks are used at the base station and mobile device (to reflect different antenna numbers and form factors). The larger the codebook, the finer the control of the resulting antenna beam. The base station and mobile device scan through their respective codebooks to find the best pair for transmission/reception. A common predefined set of BF codes was defined by the IEEE 802.15.3c standard. The phase weights W applied to a uniform rectangular array (URA) comprising $N_x \times N_z$ elements is given by Eq. 1 [1, Ch. 9], where the function fix{} returns the biggest integer smaller than or equal to its argument, $k_i = 0{:}K_i - 1$ represents the index of code k_i in dimension $i = x, z$, K_i is the length of the codebook with $K_i \geq N_i$ in dimension i, $n_i = 0{:}N_i - 1$, and $N_x \times N_z$ is the size of the array. The total codebook size is $K_x \times K_z$. For a maximum fluctuation of 1 dB in achieved gain across all pointing angles covered by the codebook, the parameter K_i must be equal to $2 \times N_i$, thus generating a large codebook with size of $4 \times N_x \times N_z$. The IEEE 802.15.3c codebook was designed for indoor applications and considers only relatively small numbers of antenna elements (< 100). While the 802.15.3c protocol specifies a hierarchical search over progressively narrower beams, the overhead still increases for larger numbers of antenna elements. Furthermore, fine phase resolution becomes necessary, which is not necessarily practical in consumer applications where cost is a key factor.

Beamforming Weight Optimization Based on AoD Estimation — Because compact arrays are likely to be used in mmWave mobile devices, the paths of the channel can be distinguished based on different AoDs, which can be estimated using classical array processing algorithms such as MUSIC, ESPRIT, and their derivatives [1, 7]. Recently, to exploit the spatial characteristics of the mmWave channel, new techniques based on compressive sensing (CS) have been proposed [17]. CS is a signal processing technique for efficiently acquiring and reconstructing a signal by taking advantage of the signal's sparseness or compressibility, and solving underdetermined linear systems. It is reported in [14, 15] that the average number of multipath spatial clusters reaching the mobile receiver is small — only 2.4 on average in NLOS channels. Therefore, CS can be efficiently used to estimate the sparse spatial propagation channel and identify the AoD. This method requires only a fraction of the codebook-based overhead and provides an accurate estimate of the AoD [17]. Given knowledge of the user's spatial channel, analog BF can be applied through the application of antenna phase weights. AoD estimation can be exploited in hybrid BF architectures to estimate the precoding (for SM) and antenna BF weights [18]. This drives system capacity to operate close to a fully digital implementation while requiring only a fraction of the radio hardware.

Long-Term BF — The feedback and tracking rate can be significantly reduced by a standard procedure known as *long-term BF* [19]. Although the gains along individual multipath components can change with small-scale motion (depending on the channel bandwidth or symbol time resolution [3]), the macro-level angular directions change at a much slower timescale. This phenomenon causes the channel to be concentrated in subspaces that remain constant over longer periods. Long-term BF identifies and aligns the signals on these subspaces, thus reducing the dimensionality in the tracking algorithms. Studies based on the NYC data above have demonstrated that long-term BF to a 1D subspace in the mmWave range would result in at most 2–3 dB of loss compared to instantaneous BF [13].

CSI at the transmitter, when used appropriately, achieves the highest performance of any MIMO system. The two classic approaches for obtaining CSI at the transmitter are through channel reciprocity and feedback.

Measurement-Based MIMO Channel Analysis

The NYU WIRELESS mmWave 2D SSCM (first reported in [14]) can be used to determine MIMO performance using SM. Figure 2 shows

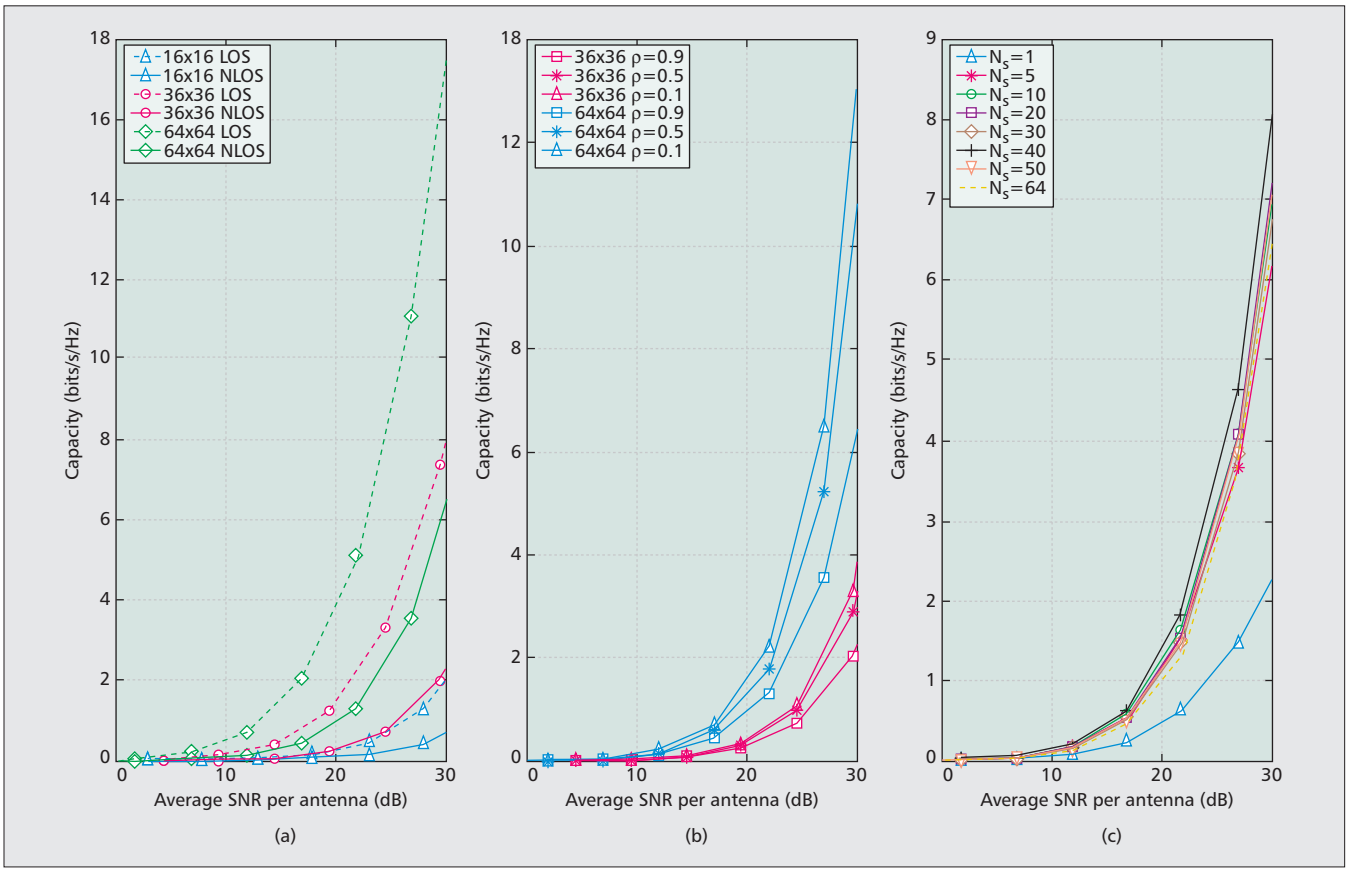


Figure 4. Channel capacity as a function of average SNR per receive antenna for a) different numbers of transmit and receive antenna elements with a correlation coefficient of 0.9 between adjacent antenna elements both at the transmitter and at the receiver; b) different values of correlation coefficient ρ between adjacent antenna elements both at the transmitter and at the receiver in NLOS environments; c) various number of transmit streams N_s with $N_t = 64$ transmit antennas and $N_r = 64$ receive antennas with $\rho = 0.9$ at 28 GHz using uniform linear arrays (ULAs) in NLOS environments.

the spatial azimuth AoD distribution in both LOS and NLOS environments derived from the omnidirectional outdoor channel model. As can be observed, AoDs in a LOS environment are typically concentrated in a narrow spatial range, while NLOS AoDs are usually more spread out, indicating the relatively high spatial degrees of freedom that can be exploited in a MIMO system at mmWave frequencies in dense urban environments.

In a MIMO system, a time-invariant (where the channel gain does not depend explicitly on time) or slow-fading (where the channel gain can be considered roughly constant over the period of use) wireless channel with N_t transmit and N_r receive uniform linear array (ULA) antennas is described by an N_r by N_t deterministic matrix $\mathbf{H}$. In the channel matrix $\mathbf{H}$, h_{ij} represents the channel gain from transmit antenna j to receive antenna i. The signals from the transmit antennas are subject to the total power constraint P. The mutual information can be expressed as

$$C = \log_2\left[\det\left(\mathbf{I} + \frac{P}{N_t N_0}\mathbf{Q}\right)\right] \text{(bits/s/Hz)} \quad (2)$$

where $\mathbf{I}$ is an $N_s \times N_s$ identity matrix, N_s is the number of transmitted spatial streams, N_0 is the total noise power, and $\mathbf{Q} = \mathbf{H}\mathbf{H}^H$ ($N_t \geq N_r$) or $\mathbf{H}^H\mathbf{H}$ ($N_t < N_r$) with $(\cdot)^{\mathrm{H}}$ denoting the conjugate transposition. While C is not strictly the channel capacity (since the source input distribution is not optimized), it is widely used as a proxy for channel capacity. Note that for closed-loop SM with eigen beamforming (EBF) where transmit precoding is used based on the right singular vectors of $\mathbf{H}$, the source distribution is optimized such that Eq. 2 leads to the correct use of the term channel capacity.

Figure 4 shows the information capacity of SM as a function of average SNR per receive antenna for different numbers of transmit and receive antenna elements at 28 GHz using ULAs. Since antennas are usually closely spaced in the ULAs, various correlation factors are assumed between different antennas. The figure shows information capacity in different environments (LOS or NLOS), with various correlation coefficients between the antenna elements in the ULAs, as well as a comparison of performance of closed-loop SM plus EBF using different numbers of transmit streams N_s, with 64 transmit antenna elements and 64 receive antenna elements. Equal power allocation is employed at all antennas at the transmitter, and the correlation coefficients between adjacent antenna elements at the transmitter and receiver are assumed to be identical. As shown in Fig. 4, the channel capacity in LOS conditions is significantly larger

than in NLOS conditions, due to the large channel gain of the dominant LOS path as compared to the rays with weaker energy in NLOS environments caused by scattering and reflections. In addition, for the same number of transmit and receive antennas, the channel capacity becomes lower when the correlation coefficient between adjacent antenna elements at the transmitter and receiver is greater, since the spatial correlation between antennas reduces the number of available eigen channels (i.e., lowers the rank) and thus degrades the total capacity from SM. Furthermore, the number of transmit streams N_s has a significant impact on the performance of the MIMO channel, and there is no monotonic relationship between N_s and channel capacity. For instance, as can be seen from Fig. 4c, for an average SNR of 20 dB and a 64 × 64 MIMO system, the capacity is maximum for N_s = 40, yet the capacity decreases when N_s either decreases or increases. This indicates that we need to search for the optimal value of N_s in order to maximize the channel capacity for a given number of transmit and receive antennas at a particular SNR operating point.

Polarization-Based Hybrid Spatial Multiplexing: Large Arrays with Analog Beamforming

For longer-distance mmWave communications at sub one watt power levels, exceeding 100 m or so, BF gain may be preferred over spatial multiplexing (especially in LOS channels) to overcome propagation path loss. BF exploits the BS array to generate the gains necessary for range extension by overcoming mmWave path loss [1, 9]. The surprising amounts of scattering and delay spread (Tables 1 and 2) yield the capacity gains from SU-MIMO SM in Fig. 4. However, as seen in Fig. 3, it is challenging to separate clustered users wholly in the spatial domain. Instead, polarization can be exploited using a hybrid BF architecture (Fig. 1) to combine the multiplexing and BF processes [12]. In this case, two I-Q baseband multiplex streams are supported (to match the two orthogonal polarizations). The nature of signal propagation in the mmWave bands is highly affected by the polarization of the transmit and receive antennas, as well as the environment [1, 20].

Single-user dual-stream SM (two radio chains) combined with orthogonal polarization and BF results in two independent transmit spatial streams, each with enhanced gain. Using the University of Bristol's Manhattan-style ray model, simulations were performed for an 8 × 8 (N_t = 64) BS antenna array where each RF chain drives a 4 × 8 array in MIMO configuration (N_s = 2). Mobiles were assumed to use a two-element dual-polar patch array (peak gain of 6.8 dBi). Users were distributed up to 80 m from the BS. As shown in Fig. 5, the theoretic capacity (based on a bandwidth of 2.16 GHz) tends toward the full channel capacity. This approach can be used to double the theoretic physical layer (PHY) throughput for a given user link. For example, adopting the IEEE 802.11ad PHY, a net peak throughput of 14 Gb/s could be achieved. It is important to note, however, that

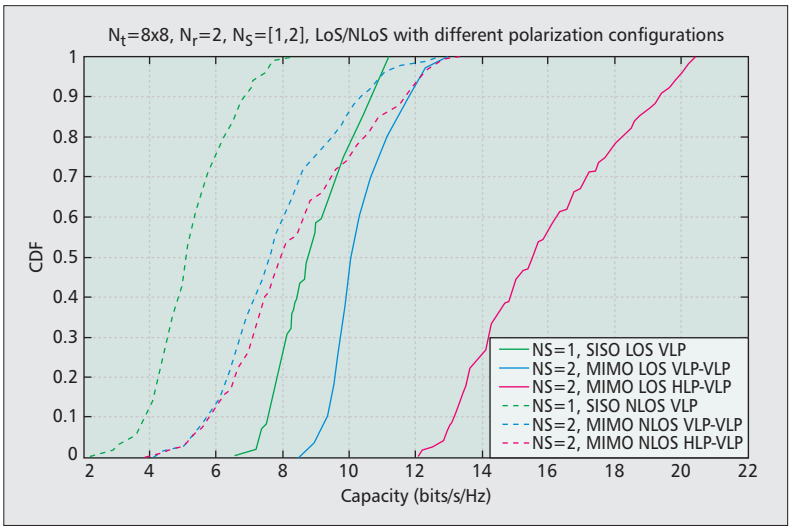


Figure 5. Channel capacity for different 2×2 single-user SM polarization schemes (LOS and NLOS configurations).

polarization diversity must be proven to work in vegetation, rain, and crowds, and more propagation measurements are needed to determine the degree of orthogonality that can be achieved. Work is currently underway at NYU WIRELESS to determine the potential of polarization diversity in real-world mmWave scenarios.

Using the same polarization on each radio chain, in this case vertical linear polarization between the transmitter and receiver (VLP-VLP), is shown to lack substantial diversity that serves to lower system performance, as shown in Fig. 5. The high gap in channel capacity when compared to cross-polarized channels is due to the high channel correlations caused by the use of the same polarization. An analysis of the eigenvalues of the VLP-VLP configuration showed undesirable high eigenvalue spreads and fast divergence vs. distance. In general, for NLOS environments the low achieved SNR leads to reduced system performance. However, for the VLP-VLP configuration, the eigenvalue spread is reduced in NLOS conditions, leading to system performance closer to that experienced using HLP-VLP.

Conclusion

In this article, we describe the fundamentals of MIMO spatial multiplexing (SM) and beamforming (BF) for future mmWave wireless communication systems, and explore when either, or both, are most likely to be useful. Measurements and realistic channel models are used to analyze and simulate results that show SM and BF could work in tandem. Among the contributions in this work, we provide extensive urban directional channel measurements at 28 and 73 GHz that show directional beams often exist with both small multipath time dispersion and low path loss, meaning that high link power can be received using BF without a complicated equalization process. We also show how mmWave systems can operate in both noise-limited and power-limited regimes, making both SM and BF potentially useful in different operating scenar-

With hybrid BF, new approaches for designing the BF weights are required as compared to lower frequency MIMO. Codebook-based beam training and AoD estimation that leverage compressive sensing techniques are promising alternatives.

ios. We present measurements that show LOS channels are often of low rank, meaning that only a couple of SM transmit streams can be supported. Polarization provides further improvement for mmWave MIMO, especially for LOS channels. While NLOS channels offer more rank, they have greater path loss, and thus will likely be best suited for exploiting both SM and BF for capacity gains. While digital BF is not practically suited to mmWave MIMO due to the large number of RF chains, as well as the expected large channel bandwidths and large numbers of antennas, we show that a generalized hybrid BF approach that exploits analog BF promises to achieve a compromise between high performance and reasonable power consumption. With hybrid BF, new approaches for designing the BF weights are required as compared to lower-frequency MIMO. Codebook-based beam training and AoD estimation that leverage compressive sensing techniques are promising alternatives.

While this article emphasizes SU-MIMO communication from the transmitter perspective, there are many indications that MU-MIMO also has great potential in mmWave networks. The presence of low-rank channels in some instances indicates that multiplexing gains can be achieved by using MU-MIMO to send data to multiple users using the hybrid BF architecture.

Acknowledgment

The authors wish to thank Mr. Mathew Samimi and Mr. Chris Slezak of NYU WIRELESS, Mr. Djamal Berraki and Dr. Simon Armour at the University of Bristol, and Prof. Gerhard P. Fettweis at Technische Universität Dresden for their contributions to this article. The authors also thank the NYU WIRELESS and WNCG industrial affiliate sponsors for supporting this work.

References

[1] T. S. Rappaport *et al*., *Millimeter Wave Wireless Communications*, 2015, Pearson/Prentice Hall.
[2] Z. Pi and F. Khan, "An Introduction to Millimeter-Wave Mobile Broadband Systems," *IEEE Commun. Mag*., vol. 49, no. 6, June 2011, pp. 101–07.
[3] T. S. Rappaport *et al*., "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, 2013, pp. 335–49.
[4] T. S. Rappaport, J. N. Murdock, and F. Gutierrez, "State of the Art in 60-GHz Integrated Circuits and Systems for Wireless Communications," *Proc. IEEE*, vol. 99, no. 8, Aug. 2011, pp. 1390–1436.
[5] D. J. Love *et al*., "An Overview of Limited Feedback in Wireless Communication Systems," *IEEE JSAC*, Special Issue on Exploiting Limited Feedback in Tomorrow's Wireless Communication Networks, vol. 26, no. 8, Oct. 2008, pp. 1341–65.
[6] G. Li *et al.*, "MIMO Techniques in WiMAX and LTE: A Feature Overview," *IEEE Commun. Mag*., May 2010, pp. 86–92.
[7] J. C. Liberti and T. S. Rappaport, *Smart Antennas for Wireless Communications*, Pearson/Prentice-Hall, 1999.
[8] R. B. Ertel *et al*., "Overview of Spatial Channel Models for Antenna Array Communication Systems," *IEEE Pers. Commun*., vol. 5, Feb. 1998, pp. 10–22.
[9] S. Sun and T. S. Rappaport, "Wideband mmWave Channels: Implications for Design and Implementation of Adaptive Beam Antennas," *Proc. 2014 IEEE MTT-S Int'l. Microwave Symp.*, June 2014, pp. 1–4.
[10] T. Bai, A. Alkhateeb, and R. W. Heath, Jr., "Coverage and Capacity of Millimeter Wave Cellular Networks," *IEEE Commun. Mag*., vol. 52, no. 9, 2014, pp. 70–77.
[11] W. Hong *et al*., "Design and Analysis of a Low-Profile 28 GHz Beam Steering Antenna Solution for Future 5G Cellular Applications," *Proc. 2014 IEEE MTT-S Int'l. Microwave Symp.*, June 2014, pp. 1–4.
[12] A. Ghosh *et al*., "Millimeter-Wave Enhanced Local Area Systems: A High-Data-Rate Approach for Future Wireless Networks," *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1152–63.
[13] M. Akdeniz *et al.*, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *IEEE JSAC*, vol. 4, issue 6, June 2014, pp. 1164–79.
[14] M. K. Samimi and T. S. Rappaport, "Ultra-Wideband Statistical Channel Model for Non Line of Sight Millimeter-Wave Urban Channels," *Proc. IEEE GLOBECOM*, Dec. 2014.
[15] M. Samimi *et al.*, "28 GHz Angle of Arrival and Angle of Departure Analysis for Outdoor Cellular Communications Using Steerable Beam Antennas in New York City," *Proc. IEEE VTC-Spring*, 2013 IEEE 77th , 2–5 June 2013, pp.1, 6.
[16] G. R. MacCartney and T. S. Rappaport, "73 GHz Millimeter Wave Propagation Measurements for Outdoor Urban Mobile and Backhaul Communications in New York City," *Proc. 2014 IEEE ICC*, June 2014, pp. 4862–67.
[17] D. E. Berraki, S. M. D. Armour, and A. R. Nix, "Application of Compressive Sensing in Sparse Spatial Channel Recovery for Beamforming in mmWave Outdoor Systems," *Proc. IEEE WCNC '14*, Apr. 6–9, 2014.
[18] O. El Ayach *et al.*, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," *IEEE Trans. Wireless Commun*., vol. 13, no. 3, Mar. 2014, pp. 1499–1513.
[19] A. Lozano, "Long-Term Transmit Beamforming for Wireless Multi-Casting," *Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Processing*, Apr. 2007, pp. III-417–III-420.
[20] A. Maltsev *et al.*, "Impact of Polarization Characteristics on 60-GHz Indoor Radio Communication Systems," *IEEE Antennas and Wireless Propagation Letters*, vol. 9, 2010, pp. 413–16.

Biographies

Shu Sun (ss7152@nyu.edu) received her B.S. degree in applied physics from Shanghai Jiao Tong University, China, in 2012. She is currently working toward a Ph.D. degree in electrical engineering at Polytechnic School of Engineering, New York University (NYU), Brooklyn, and is doing research at the NYU WIRELESS research center. She has authored or co-authored more than 10 technical papers in the field of wireless communications. Her current research interests include mmWave mobile communications and the analysis of MIMO systems for mmWave channels.

Theodore S. Rappaport (tsr@nyu.edu) received his B.S., M.S., and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, Indiana, in 1982, 1984, and 1987, respectively. He is an Outstanding Electrical and Computer Engineering Alumnus and Distinguished Engineering Alumnus from his alma mater. He holds the David Lee/Ernst Weber Chair in Electrical and Computer Engineering at Polytechnic School of Engineering, NYU, and is a professor of computer science and professor of radiology at NYU. In 2012, he founded NYU WIRELESS, a multidisciplinary research center involving NYU's engineering, computer science, and medical schools. Earlier in his career, he founded the Wireless Networking and Communications Group (WNCG), University of Texas at Austin. He has authored or co-authored more than 200 technical papers, over 100 U.S. and international patents, and several best-selling technical books. He was elected to the Board of Governors of the IEEE Communications Society (ComSoc) in 2006, and was elected to the Board of Governors of the IEEE Vehicular Technology Society (VTS) in 2008 and 2011.

Robert W. Heath, Jr. (rheath@utexas.edu) received his B.S. and M.S. degrees in electrical engineering from the University of Virginia, Charlottesville, in 1996 and 1997, respectively, and his Ph.D. in electrical engineering from Stanford University, California, in 2002. Since January 2002, he has been with the Department of Electrical and Computer Engineering, University of Texas at Austin, where he is currently a professor and director of the Wireless Networking and Communications Group. His research interests include several aspects of MIMO communication: limited feedback techniques, multihop networking, multiuser MIMO, antenna design, and scheduling algorithms, as well as 60 GHz communication techniques and multimedia signal processing. He is a Registered Professional Engineer in the State of Texas. He is the recipient of the David and Doris Lybarger Endowed Faculty Fellowship in Engineering. He has been

an Editor for *IEEE Transactions on Communications* and an Associate Editor for *IEEE Transactions on Vehicular Technology*. He is a member of the Signal Processing for Communications Technical Committee of the IEEE Signal Processing Society and the Vice Chair of the IEEE Communications Society Communications Technical Theory Committee.

ANDREW NIX (Andy.Nix@bristol.ac.uk) received his B.Eng. and Ph.D. degrees from the University of Bristol in 1989 and 1993, respectively. He joined the Centre for Communications Research as a member of lecturing staff in December 1992, and is currently a professor of wireless communication systems, head of the Communication Systems & Networks Group, and Dean of the Faculty of Engineering. His research interests are focused around the challenges associated with ubiquitous, mass market, low-cost mobile telecommunication systems. This includes the development of innovative spectral and power-efficient digital wireless technologies. Over a period of 22 years he has supervised 51 Ph.D. students to completion and published in excess of 400 refereed journal and international conference papers.

SUNDEEP RANGAN (srangan@nyu.edu) received his B.A.Sc. at the University of Waterloo, Canada, and his M.Sc. and Ph.D. at the University of California, Berkeley, all in electrical engineering. He has held postdoctoral appointments at the University of Michigan, Ann Arbor and Bell Labs. In 2000, he co-founded (with four others) Flarion Technologies, a spin-off of Bell Labs, which developed Flash OFDM, the first cellular OFDM data system. Flarion grew to over 150 employees with trials worldwide. In 2006, Flarion was acquired by Qualcomm Technologies. He was a director of engineering at Qualcomm involved in OFDM infrastructure products. He joined the ECE Department at NYU Polytechnic School of Engineering in 2010. His research interests are in wireless communications, signal processing, information theory, and control theory.

MIMO Precoding and Combining Solutions for Millimeter-Wave Systems

Ahmed Alkhateeb, Jianhua Mo, Nuria González-Prelcic, and Robert W. Heath Jr.

ABSTRACT

Millimeter-wave communication is one way to alleviate the spectrum gridlock at lower frequencies while simultaneously providing high-bandwidth communication channels. MmWave makes use of MIMO through large antenna arrays at both the base station and the mobile station to provide sufficient received signal power. This article explains how beamforming and precoding are different in MIMO mmWave systems than in their lower-frequency counterparts, due to different hardware constraints and channel characteristics. Two potential architectures are reviewed: hybrid analog/digital precoding/combining and combining with low-resolution analog-to-digital converters. The potential gains and design challenges for these strategies are discussed, and future research directions are highlighted.

INTRODUCTION

Communication over millimeter-wave (mmWave) frequencies is defining a new era of wireless communication, and most recently cellular systems [1, 2]. Recent studies show that the combination of high-bandwidth channels, network densification, and large antenna arrays at both the base station and mobile users can provide coverage comparable to conventional lower-frequency networks [3] but with much higher data rates. Reaping the gains offered by mmWave, however, requires multiple-input multiple-output (MIMO) signal processing, which leverages the higher aperture created by the antenna arrays in a way that respects the hardware design challenges in mmWave circuits. Commercial mmWave systems like IEEE 802.11ad use single-stream MIMO transmission. This article explores potential architectures for using more sophisticated MIMO precoding and combining at mmWave.

MIMO precoding/combining in mmWave systems is generally different than precoding at lower frequencies, for example, the UHF frequencies used in current cellular systems. One reason is that **hardware constraints are different**: while the small wavelength of mmWave signals allows a large number of antennas to be packed into a small form factor, the high cost and power consumption of some mixed signal components, like high-resolution analog-to-digital converters (ADCs), makes it difficult to dedicate a separate complete radio frequency (RF) chain with these components for each antenna [1]. This makes the conventional architecture in current cellular systems — where precoding and combining are performed entirely in the digital baseband — infeasible. A second difference is that MIMO systems in mmWave will make use of a **large number of antennas**. This impacts the complexity of key signal processing functions like channel estimation, precoding, combining, and equalization. Moreover, **mmWave propagation characteristics are different**, so that the MIMO channel is not as "rich" at mmWave. For example, measurements show that the mmWave channel is sparse in the angular domain [4], which can be leveraged to realize efficient precoding/combining algorithms [5]. Finally, mmWave communication channels will use a **large bandwidth**, meaning that broadband channel equalization will still be required. Because of the hardware constraints, the large number of antennas, the different channel conditions, and the larger channel bandwidth, new MIMO transceiver architectures are needed for mmWave systems.

In this article, we present two potential mmWave MIMO transceiver architectures inspired by the hardware constraints while still providing high data rates.

The first solution is hybrid analog/digital precoding (and combining) in which the required precoding and beamforming are divided between the analog and digital domains. The digital precoding layer adds more freedom for the precoding design problem compared to a pure analog beamforming solution. This enables hybrid precoding to achieve better performance, especially for multi-stream and multi-user transmission. The algorithms in [5, 6, 15] leverage mmWave channel characteristics, such as channel sparsity, to realize low-complexity but highly efficient hybrid precoding and channel estimation solutions.

The second solution is the use of low-resolution ADCs to reduce power consumption at the receiver. Note that the digital-to-analog converters (DACs) at the transmitter consume less power; therefore, we focus only on receiver tech-

Ahmed Alkhateeb, Jianhua Mo, and Robert W. Heath, Jr. are with the University of Texas at Austin.

Nuria González-Pelcic is with Universidade de Vigo.

0163-6804/14/$25.00 © 2014 IEEE

niques. Low-resolution ADCs can be implemented with simpler circuits, consume less power, and at low signal-to-noise ratio (SNR) incur only a little rate loss compared to high-resolution quantization [7]. This approach is seen as an alternative to the hybrid beamforming paradigm where, instead of combining mostly in analog and using high-resolution sampling, all combining is performed in digital based on very coarse quantization. Both approaches can coexist in the same system. For instance, in a cellular system, a power-limited mobile station might adopt combining with low-resolution ADCs on the downlink, while hybrid precoding/combining might be used for the uplink or the backhaul between base stations.

The goal of this article is to review the emerging area of mmWave MIMO precoding and receiver design, with an emphasis on the communication and signal processing aspects. First, we explain how mmWave precoding is different than lower-frequency techniques. Then we discuss state-of-the-art analog beamforming solutions that were developed primarily for indoor mmWave communication and explain how they are not sufficient for cellular communication. Next we introduce two mmWave precoding/combining solutions: hybrid analog/digital precoding/combining and combining with low-resolutions ADCs. We compare the performance of the two solutions in simulations and draw conclusions about how they should be employed in mmWave systems. Finally, we highlight future research directions, including performance improvements and possible extensions.

MmWave Precoding Is Different

Different Hardware Constraints

Operating in mmWave frequencies with wide bandwidths imposes additional hardware constraints that impact current transceiver architectures. Therefore, changes need to be made on these architectures to meet power budgets and reduce implementation costs. This in turn imposes new and different constraints on the mmWave precoding design problems. For instance, mixed-signal devices, like high-bandwidth high-resolution ADCs, are expensive and power-hungry. Furthermore, the baseband digital processing complexity grows with the number of ADCs. Hence, performing some beamforming or combining in analog is attractive. There are different constraints; for example, the phase shifters are subject to other hardware constraints: the angle is quantized, and the amplitude is fixed. As a result, precoding solutions need to be developed for the transceiver architectures that take hardware limitations into consideration.

Different Antenna Scales

The smaller antenna aperture at high frequencies captures less power. Hence, large antenna arrays need to be deployed at both the transmitter and receiver to provide sufficient beamforming gains and received power. For example, in future mmWave cellular networks, it seems plausible (depending on the frequency) to have 256 antennas at the base station and 16 antennas at the mobile station. However, this means that both the transmitter and receiver are required to design very large precoding and combining matrices associated with large-scale MIMO systems. The large scale also increases the complexity and overheads associated with traditional precoding and channel estimation algorithms. For example, traditional MIMO channel estimation techniques generally depend on estimating the entries of the channel matrices, which will require high training overhead in large MIMO systems. This illustrates the need to develop low-complexity precoding, channel training, and estimation algorithms for mmWave systems.

Different Channel Characteristics

MmWave has different propagation characteristics in potential bands that may be used for cellular compared to lower-frequency channels [4, 8]. For example, mmWave channels are expected to be sparse in the angular domain meaning only a few scattering clusters. Furthermore, line-of-sight (LOS) and non-line-of-sight (NLOS) signals have different path loss laws: LOS transmission is similar to free-space signals, while NLOS signals are much weaker and susceptible to environments. Compared to lower-frequency systems, there is typically less delay spread at mmWave. As many material surfaces are rough at small mmWave frequencies, mmWave signal propagation may experience larger angle spreads. The different characteristics of mmWave channels should be considered in the design of any mmWave system. For example, the system should be robust enough that it works in both LOS and NLOS. Structure in the channel (e.g., sparsity) can be exploited to reduce complexity and training overhead [5, 6]. Further work on channel models is still needed to better characterize the available sparsity, include different mobility settings, and account for blockage effects.

Different Communication Channel Bandwidth

The main motivation for shifting cellular communication to mmWave bands is to make use of the large bandwidth available at high frequencies. This enables users to be assigned a large communication channel bandwidth, and to send data at very high rates. The large communication channel bandwidth, though, impacts the mmWave cellular system operation and precoding transceiver architectures. First, large channel bandwidth results in high noise power and low received SNR before beamforming design. This makes it challenging to implement functions like random access, channel estimation, and beam training. Furthermore, broadband channels coupled with delay spread mean that equalization will likely be required at the receiver. However, the hardware constraints make it very difficult to perform the required equalization processing entirely in the baseband as done in conventional cellular systems. New algorithms and architectures are needed to operate in mmWave broadband channels, accounting for the large arrays, hardware constraints, and channel characteristics.

Operating in mmWave frequencies with wide bandwidths imposes additional hardware constraints that impact current transceiver architectures. Therefore, changes need to be made on these architectures to meet power budgets and reduce implementation costs.

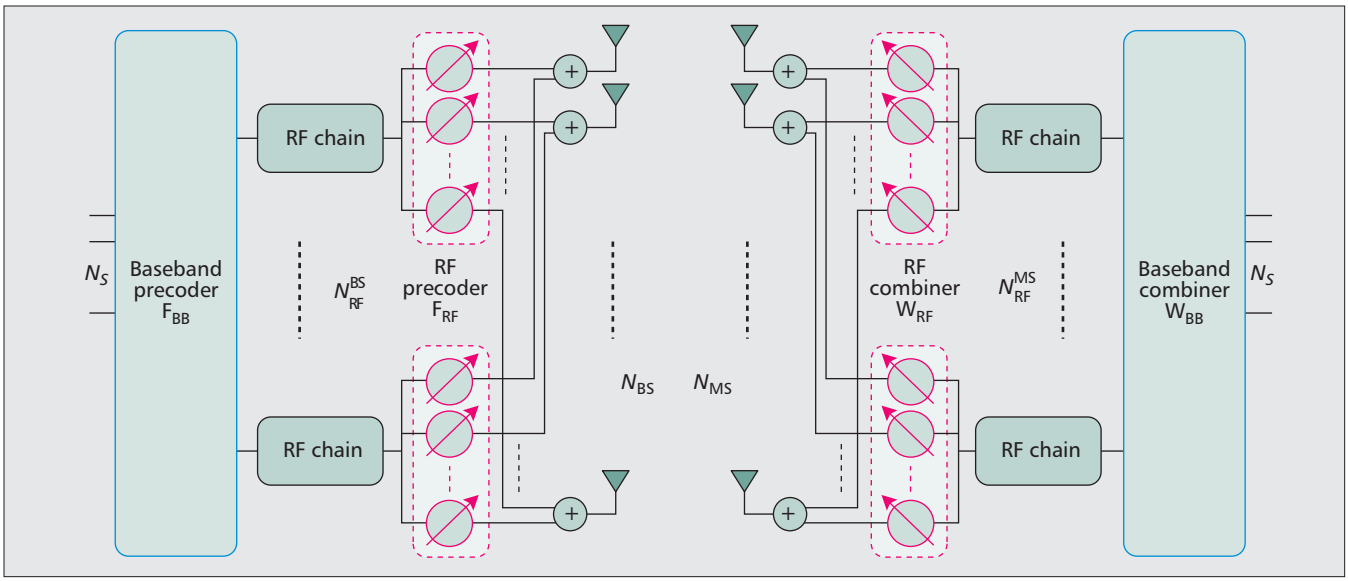


Figure 1. Hybrid analog/digital precoding and combining architecture. In this model, the number of antennas at the base station and mobile station, N_{BS}, N_{MS}, are much larger than the number of RF chains N_{RF}^{BS} and N_{RF}^{MS}, respectively. The precoding and combining processing is divided between the analog and digital domains.

Analog Beamforming

In conventional cellular systems like Third Generation Partnership Project (3GPP) Long Term Evolution (LTE), precoding and combining are performed in baseband using digital signal processing. This digital precoding allows better control over the entries of the precoding matrices, which in turn facilitates the implementation of sophisticated single-user, multi-user, and multi-cell precoding algorithms. Operating in the digital domain also permits precoding and combining to be performed in conjunction with equalization, for example in the frequency domain with orthogonal frequency-division multiple access (OFDMA). The hardware constraints discussed in the previous section, however, mean that digital baseband solutions are generally infeasible in the near term for mmWave. An immediate solution to overcome the limitation on the number of complete RF chains is to perform beamforming in analog (at RF or some intermediate frequency) using networks of phase shifters [1]. The weights of the phase shifters are designed to shape and steer the transmit and receive beams along the dominant propagation directions.

Analog beamforming/combining is the de facto approach for indoor mmWave systems. Building the beamforming vectors requires multiple phases of beam training to perform an iterative and joint design of the weights of the phase shifter networks at each node in the network in each direction. In IEEE 802.11ad, for example, these phases are sector level sweep (determining the best sector), beam refinement (sharpening the beam), and beam tracking (adjusting the beam over time). In IEEE 802.15.3c wireless personal area networks, a binary search beam training algorithm is used to progressively refine and sharpen the beams using a layered multi-resolution beamforming codebook [9].

The potential of analog beamforming is limited by the availability of only quantized phase shifters and the constraints on the amplitudes of the phase shifters. These constraints make it hard to form multiple beams, finely tune the sidelobes, or steer nulls. Furthermore, analog beamforming/combining algorithms are limited to single-stream transmission; their extension to multi-stream or multi-user cases are not straightforward.

Hybrid Analog/Digital Precoding and Combining

Digital precoding and combining allow for advanced transmission strategies, but with high complexity and power consumption in mmWave systems. Analog beamforming is relatively simple, but only supports single-stream transmission. One compromise on the performance/complexity trade-off is hybrid analog/digital precoding and combining [5, 6, 10, 15]. In hybrid precoding, the precoder processing is divided between analog and digital domains, as shown in Fig. 1.

In Fig. 1, the BS has N_{BS} antennas and N_{RF}^{BS} RF chains, and the mobile station has N_{MS} antennas and N_{RF}^{MS} RF chains such that $N_{BS} > N_{RF}^{BS}$ and $N_{MS} > N_{RF}^{MS}$. The precoding processing is divided between the analog and digital baseband precoding matrices $\mathbf{F}_{RF}$, $\mathbf{F}_{BB}$, and the combining at the mobile user is done using the analog and baseband combining matrices $\mathbf{W}_{RF}$, $\mathbf{W}_{BB}$. If $\mathbf{H}$ denotes the channel matrix, $\mathbf{s}$ represents the transmitted signal vector, and $\mathbf{n}$ represents the received noise vector, the received signal after combining is written as

$$\mathbf{y} = \mathbf{W}_{BB}^{*}\mathbf{W}_{RF}^{*}\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} + \mathbf{W}_{BB}^{*}\mathbf{W}_{RF}^{*}\mathbf{n}. \quad (1)$$

The difference between the received signal in Eq. 1 and the typical MIMO signal model is the product of precoding and combining matrices,

each implemented in a different domain and with different structural constraints.

Advantages and Limitations

Hybrid precoding provides a compromise between hardware complexity and performance gain. The number of complete RF chains required in hybrid precoding is much lower than the number of antennas, which makes it cost and power efficient. In [5, 6], hybrid precoders were designed and shown to achieve near-optimal data rates compared to digital unconstrained solutions. Thanks to the additional digital layer, hybrid precoding has more freedom in designing precoding matrices than analog beamforming. This allows hybrid precoding to perform more complicated processing, and to support multi-stream multiplexing and multi-user transmission. Moreover, this additional digital layer enables mmWave systems to operate more robustly with broadband channels, for example, to perform frequency domain space-time equalization.

Compared to fully digital baseband solutions, the performance of hybrid precoding/combining is limited by the number of RF chains. For example, the multiplexing gain of the link (the number of data streams that can be supported) is upper bounded by the minimum of the number of RF chains at the base station and mobile station. In mmWave systems, however, the channels are expected to be sparse in the angular domain, which can be exploited to reduce the performance gap between digital and hybrid precoding. Therefore, the rank of the channel matrix is less than or equal to the number of significant paths, and it can be shown that with the number of RF chains equal to the channel rank, the performance of hybrid precoding can be the same as digital precoding in single-user mmWave systems assuming the availability of unquantized phase shifters [5, references therein].

Design Challenges

The following points highlight some of the main challenges in the hybrid precoding framework for mmWave systems and discuss some related research that tackles these challenges.

Low-Complexity Precoding/Combining Designs — Maximizing throughput in single-user mmWave channels with hybrid precoding is challenging. The main difficulty comes from the coupling between analog and digital precoders, which imposes new and different constraints on the precoder design problem. For example, the design of the hybrid precoders in Eq. 1 requires designing the matrix $\mathbf{F}_{RF}\mathbf{F}_{BB}$, which is a product of two precoding matrices. Furthermore, the constant modulus constraint on the phase shifters requires the entries of the RF precoding matrix $\mathbf{F}_{RF}$ to have equal norm. This allows the columns of this matrix to be selected only from a finite set of possible RF beamforming vectors. In summary, as explained in [5], finding the precoding matrices is equivalent to solving an optimization problem with non-convex feasibility constraints, which does not have a general known solution. Hence, only approximations to the real optimization problem can be solved, so sub-optimal but low-complexity hybrid precoding/combining designs are desirable.

In this context, a first low-complexity sub-optimal design was proposed in [5]. In this work it is assumed that the mmWave channel is sparse in the angular domain: there may be only a few angles of arrival/angles of departure (AoAs/AoDs). With this assumption, the hybrid precoding/combining design problem can be formulated as a sparse approximation problem, and a variant of the matching pursuit algorithm was developed to efficiently design the analog/digital precoding and combining matrices. This design illustrated that hybrid precoding can effectively achieve performance gains comparable to digital baseband solutions while requiring much less hardware complexity, which makes it a promising precoding solution for mmWave systems. In Fig. 2, we compare the performance of hybrid analogand digital precoding/combining solutions in mmWave systems with analog and unconstrained digital solutions. The proposed hybrid precoder/combiner achieves spectral efficiencies that are very close to those achieved by the optimal unconstrained solution in the given parameters.

Channel Estimation with Hybrid Precoding — Constructing the precoding and combining matrices in the most straightforward way requires a channel estimate, which is difficult in mmWave systems. First, the channel matrix is large due to the use of large arrays. Therefore, using traditional channel estimation techniques that estimate the entries of the channel matrix requires a lot of training overhead. Second, the large mmWave communication channel bandwidth increases noise power and makes the received SNR very low before beamforming design. Third, in traditional baseband processing, there is a direct access to the entries of the channel matrix. In hybrid precoding, however, the channel seen in the baseband is through the lens of the RF precoding and combining, which further complicates the channel estimation processing and the training signal design. Building hybrid precoding/combining matrices, especially the RF beamsteering vectors, may require knowledge of the array geometry, which may not be available due to, for example, blockage of antennas on the mobile station by fingers.

In [6], we leveraged mmWave channel characteristics to realize efficient training and estimation algorithms using hybrid precoding/combining architectures. Thanks to the sparse nature of mmWave channels in the angular domain, the channel matrix can be completely defined in terms of a small number of parameters, namely: the AoA/AoD and path gain of each of the few channel paths. Estimating a mmWave channel is then equivalent to estimating the parameters of this channel. Leveraging this observation, [6] formulates the mmWave channel estimation problem as a compressed sensing problem for which the tools developed in the compressed sensing framework can be used to estimate the defining parameters, which helps reduce the required training overhead. The sparse formulation also provides an efficient way to estimate multi-path mmWave channels by

Building hybrid precoding/combining matrices, especially the RF beamsteering vectors, may require knowledge of the array geometry, which may not be available due to, for example, blockage of antennas on the mobile station by fingers.

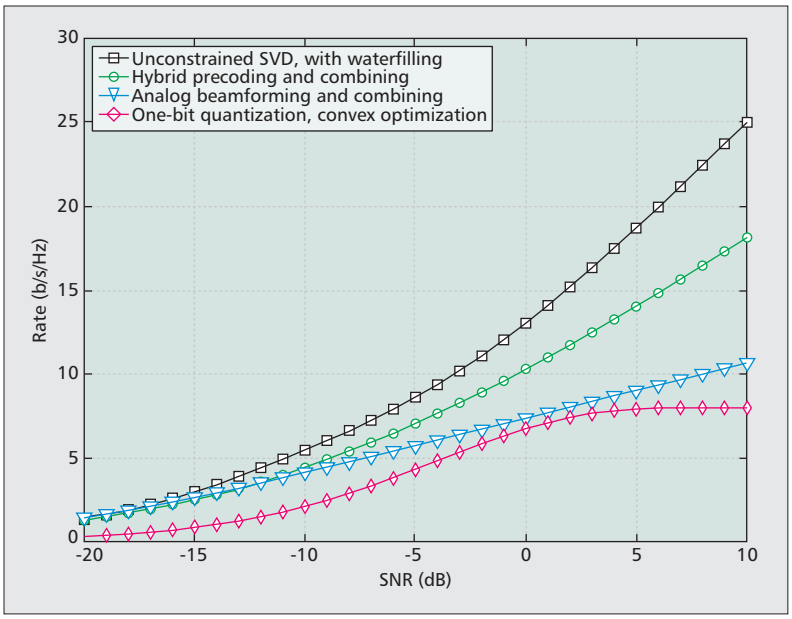


Figure 2. Spectral efficiency comparison of different precoding strategies. The figure compares the performance of the proposed hybrid analog/digital precoding/combining and 1-bit ADCs combining solutions with the optimal unconstrained SVD precoding and analog beamforming/combining.

adaptively removing the contribution of each channel path and estimating the other paths. To tackle the low-SNR problem, the base station and mobile user employ training beamforming and combining vectors in the channel estimation phase. To reduce the training overhead, a multi-resolution codebook for these training vectors was designed in [6] using hybrid analog/digital precoders. Simulation results show that multi-path mmWave channels can be estimated efficiently using adaptive compressed sensing tools while requiring relatively small training overhead compared with mmWave channel matrix dimensions. This work, however, assumes arbitrary but known array geometries for both the transmitter and receiver. Developing robust mmWave channel estimation algorithms with random and unknown antenna array geometries is an interesting open problem.

Hybrid Precoding/Combining with Limited Feedback — mmWave systems are expected to operate in a bidirectional fashion with both the base station and mobile station alternately acting as a transmitter. Consequently, it is reasonable to exploit the presence of a feedback link to aid in establishing the forward link. The feedback link has to be limited (low rate) for two reasons. First, the dimensionality of the channel and precoding matrices are quite large. Second, prior to establishing the beamforming in both directions, the data rate is generally low and achieved using either quasi-omni beam patterns and spreading or an overlaid lower frequency communication channel. The conventional approach of implementing limited feedback is a codebook-based approach. Previous limited feedback codebook designs cannot be used, due to the joint dependence between analog and digital precoders and the hardware constraints on the analog precoders.

In [5], the single-user hybrid precoding design problem for mmWave systems was considered. The performance of the developed precoders was evaluated under a limited feedback assumption. The sparse nature of mmWave channels in the angular domain was leveraged to design an efficient codebook with a small size for the RF beamforming vectors. The RF beamforming codebook consisted of beamsteering vectors in quantized directions. The digital precoding matrices were quantized on the Grassmann manifold using Lloyd's algorithm. Despite their good performance, the codebooks in [5] did not consider the relation between analog and digital precoders, which can potentially be exploited for further performance improvements.

Combining with Low-Resolution ADCs

In conventional MIMO receiver designs, the ADCs are expected to have high resolution (e.g., 6 or more bits) and act as transparent waveform preservers. In mmWave systems, the sampling rate of the ADCs scales up with the larger bandwidth. Unfortunately, high-speed (e.g., > 1 GSample/s) high-resolution ADCs are costly and power-hungry [11]. In an ideal b-bit ADC with flash architecture, there are $2^b - 1$ comparators; thus, the power consumption grows exponentially with the resolution. At present, commercially available ADCs with high speed and high resolution consume on the order of several Watts of power. Therefore, the high power assumption of high-resolution ADCs at the receiver may be a bottleneck for mmWave MIMO systems, especially on battery powered mobile stations. Possible solutions are to reduce either the sampling rate or the quantization resolution of ADCs, or both. Reducing the sampling rate can be performed with different ADC structures, for example the time-interleaved ADC (TI-ADC), where a number of low-speed high-resolution ADCs operate in parallel. The main drawback of the TI-ADC is the mismatch among the sub-ADCs in gain, timing, and voltage offset, which can cause error floors in receiver performance, although it is possible to compensate the mismatch at the price of higher complexity of the receiver. An alternative solution is to live with low-resolution ADCs (1–3 bits), which reduces power consumption and cost. In this article, we focus on the extreme case when 1-bit ADCs are used. As explained in this section, the 1-bit resolution solution has ramifications on the entire transceiver design.

Advantages and Limitations

Figure 3 illustrates a receiver structure where a 1-bit ADC is used for each in-phase and quadrature baseband received signal. The main advantage of this architecture is that the ADC can be implemented by a single comparator, which results in very low power consumption. The architecture also simplifies other aspects of circuit complexity; for example, automatic gain control (AGC) may not be required. Previous

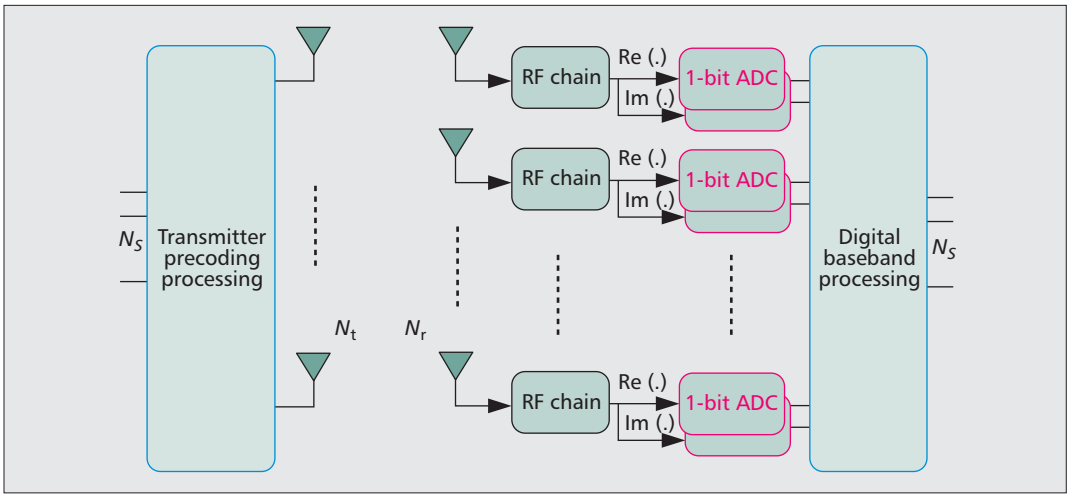


Figure 3. A system model for 1-bit ADC precoding. For each receiver antenna, there is one RF chain and two 1-bit ADCs. There are many possible choices at the transmitter, including hybrid precoding, fully digital precoding, or analog beamforming.

work has shown that ADCs with low resolution can achieve much of the unquantized capacity in the low and medium SNR regimes [7].

There are a few disadvantages associated with the 1-bit ADC model. Compared to the receiver for hybrid combining, more RF chains are required in receivers with low-resolution ADCs. The theoretical capacity that can be achieved is also limited. For example, in a channel with N_r receiver antennas, the channel capacity has an upper bound of $2N_r$ b/s/Hz since there are only $2N_r$ bits available at the output of the ADCs. Therefore, compared to high-resolution quantization, coarse quantization incurs a performance degradation in the high SNR region.

Design Challenges

Achieving the gains promised by 1-bit ADCs combining requires overcoming a number of design challenges, as explained in the following points. In general, the nonlinearity of the quantization operation makes the analysis difficult. Approximations or iterative optimizations are usually used to sidestep this difficulty.

Capacity Analysis with 1-Bit ADCs — With 1-bit quantization, the outputs of the ADCs are discrete and finite. As a result, the optimal transmitted signals have been proven to be discrete and finite [7]. This is in contrast with unquantized MIMO systems where the optimal transmitted signals are Gaussian. Finding the channel capacity and the capacity optimizing transmit signal distribution is a challenging problem.

In [12], we considered the channel capacity of the flat-fading narrowband MIMO channel with 1-bit ADCs shown in Fig. 3. In the simple case of the SIMO channel, the optimal transmitted signals were found by a numerical algorithm. It was proven that the rate scales with $\log_2(N_r)$ at high SNR. Furthermore, bounds for the high SNR capacity of the MIMO channel with 1-bit quantization were provided. These results for the general MIMO channels can also be applied to sparse mmWave channels, where it was found that capacity is mainly limited by the number of paths.

Channel Estimation with 1-Bit ADCs — Most capacity analysis with low-resolution ADCs is based on the critical assumption that the transmitter and receiver have complete and perfect channel knowledge. There is a channel estimation error, however, in practical systems, and this error could be quite high when using 1-bit ADCs. It is of interest to develop accurate channel estimation techniques at the receiver or transceiver designs that avoid the need for channel estimation. An effective way to estimate the single-input single-output (SISO) channel is to use dithering to combat the severe nonlinearity of 1-bit quantization. Dithering means that a special signal is added to the received signal before quantization. Even with 1-bit quantization, it is possible to attain near infinite resolution performance at the price of longer training sequences.

For the MIMO channel, a channel estimation method using an expectation-maximization (EM) algorithm was proposed in [13] to find the maximum a posteriori probability estimate. It can be observed that above a certain SNR, in the quantized case, the estimation error increases with SNR instead of decreasing. This phenomenon is known as stochastic resonance. A simple explanation is that the 1-bit ADC is a highly nonlinear system in which noise may actually be helpful in some circumstances.

Broadband Channels with 1-Bit ADCs — MmWave systems will work in broadband channels. For example, IEEE 802.11ad specifies channels with bandwidths of up to 2.16 GHz and is designed to deal with delay spread around 10–40 ns in the indoor environment. Equalization seems to be challenging since coarse quantization of the received signal occurs after the convolutive effects of the channel. Consequently, it is of interest to develop receiver architectures where other functions are performed in analog prior to the ADC.

Achieving the gains promised by one-bit ADCs combining requires overcoming a number of design challenges as explained in the following points. In general, the nonlinearity of the quantization operation make the analysis difficult. Approximations or iterative optimizations are usually used to sidestep this difficulty.

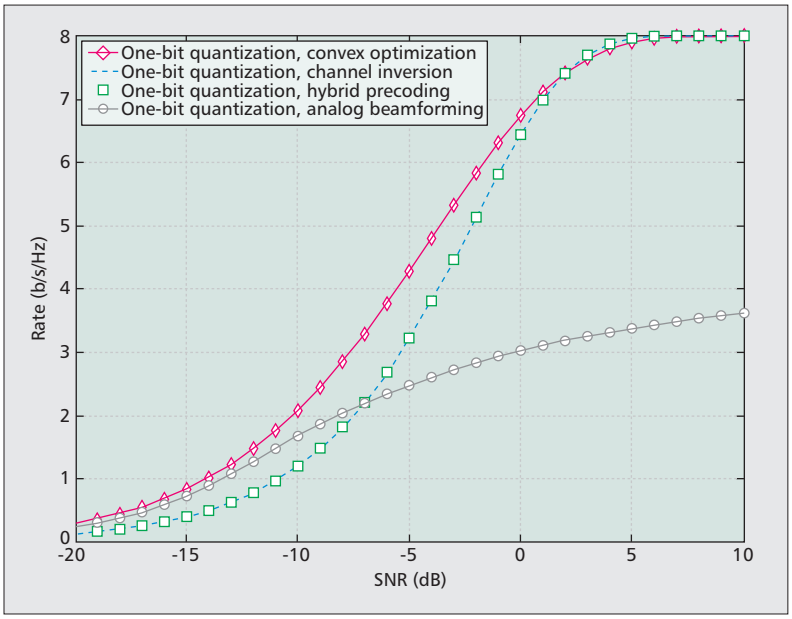


Figure 4. The achievable rates for different transmission strategies and receiver structures.

One promising approach is to revive the analog discrete Fourier transform (DFT), a topic of very early research on the DFT, and place this analog circuit prior to the ADCs [14]. With an analog DFT, the signals are orthogonalized prior to sampling, so the sampled signal will ideally not have any intersymbol interference or intercarrier interference. Other advantages include lower power than digital DFTs and smaller dynamic range.

Performance Comparisons

In this section, we illustrate the performance of the proposed multi-stream mmWave-suitable precoding/combining strategies, and compare them with traditional multi-stream baseband solutions. We adopt a geometric mmWave channel model characterized by few paths between the base station and mobile user to capture the sparse nature of the channel [4, 8]. The channel is assumed to be perfectly known at the transmitter and receiver. Both the base station and mobile user are assumed to employ antenna arrays of different numbers to provide sufficient beamforming gains.

In Fig. 2, we show the achievable rates in a mmWave channel with 64 base station antennas and 4 mobile user antennas. The channel is assumed to have four paths, and the angles of arrival and departure are uniform random variables in $[0, 2\pi]$. The figure compares four precoding schemes: hybrid [5], analog, baseband transmitter and 1-bit ADC receiver, and baseband unconstrained singular-value decomposition (SVD) precoding solutions. Four streams are assumed to be transmitted simultaneously in all cases except for analog-only beamforming where single-stream transmission is assumed. For the hybrid precoding, the base station is assumed to have eight RF chains, while the mobile user has three RF chains. The analog beamforming vectors of the hybrid precoding are taken from a quantized beamsteering codebook where the steering directions are uniformly quantized with 7 bits. Figure 2 shows that hybrid precoding can achieve spectral efficiency comparable to the SVD solution. In our prior work [5, 6], we have shown that very close rates to the SVD solution can be achieved by hybrid precoding/combining even when only partial and imperfect channel knowledge (only quantized AoAs/AoDs knowledge) exists. Since only one stream is transmitted, analog beamforming/combining has a smaller multiplexing gain compared to the hybrid scheme. The baseband transmitter with 1-bit ADCs has the worst performance. Its rate saturates in the high SNR region. Because of the large bandwidths, SNRs in mmWave will be moderate, so high SNR saturation is not necessarily a major limiting factor. This 1-bit quantization approach, however, has the least mixed-signal power consumption, which may be an acceptable trade-off for ultra-low-power implementations.

In Fig. 4, we show the achievable rates of different transmission strategies in a mmWave channel with 1-bit ADC receivers. As in Fig. 2, there are 64 transmitter antennas, 4 receiver antennas, and 4 paths in the channel. Four streams are assumed to be simultaneously transmitted. The number of RF chains with 1-bit ADCs at the mobile user is equal to the number of antennas, which is 4. First, we find that the achievable rates with 1-bit quantization are lower than the upper bound of 8 b/s/Hz. This verifies our analysis that the achievable rate with 1-bit quantization is upper bounded by $2N_r$ b/s/Hz. Second, we compare four different transmission strategies with 1-bit ADC receivers. In the case of convex optimization, the transmitted symbols are designed using a convex optimization algorithm. For the case of channel inversion, channel inversion precoding and QPSK signaling are adopted. We see that the convex optimization provides the best performance, at the expense of higher complexity. Channel inversion has the lowest complexity, and it works well in the moderate-to-high SNR region. In the case of hybrid precoding, we assume there are eight RF chains at the transmitter. The rate of the hybrid precoding is very close to channel inversion, where 64 RF chains are needed. The approach of analog beamforming where only one RF chain is required has the worst performance.

In Fig. 5, we plot the upper bound of the mmWave channel capacity with 64 transmitter antennas [12]. The receiver antennas and the number of paths are both assumed to be no more than 16. In this setup, the bound only depends on the number of paths L and the number of receiver antennas N_r. First, we see that the upper bound is mostly limited by the number of paths when L is small compared to N_r. When L is small, the upper bound almost increases linearly with L. This means that the channel capacity is limited by the number of paths when there are a lot of receiver antennas. Second, although the upper bound increases with L for fixed N_r, it saturates to $2N_r$ b/s/Hz when $L \geq N_t$. This means that in a non-sparse channel, the capacity is limited by the 1-bit quantization at the receiver.

Finally, we study the impact of hardware constraints on the performance of hybrid precoding in Fig. 6. In this figure, the spectral efficiency achieved by hybrid precoding is compared to that of the SVD unconstrained solution with different numbers of RF chains and phase shifter quantization bits. In this simulation, four streams are simultaneously transmitted, and two setups are compared. In the first one, the base station has $N_{RF}^{BS} = 8$, and mobile user has $N_{RF}^{MS} = 4$. In the second one, the base station has $N_{RF}^{BS} = 4$, and the mobile user has $N_{RF}^{MS} = 2$. The number of antennas are assumed to be $N_{BS} = 64$ and $N_{MS} = 4$, and the channel is assumed to have 4 paths. Results show that 6–7 quantization bits may be enough to achieve near-optimal performance and that very low quantization has a huge performance penalty. Also, the figure illustrates that the additional RF chains have a positive impact on the achieved spectral efficiency due to the better approximation of the digital precoders and combiners.

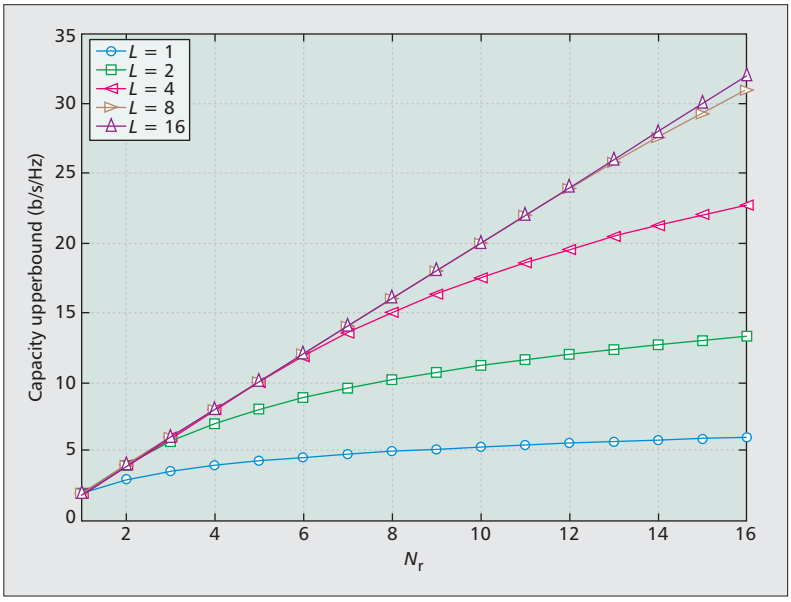


Figure 5. The upper bound of the channel capacity vs. the number of receiver antennas N_r for different s of paths L in the channel.

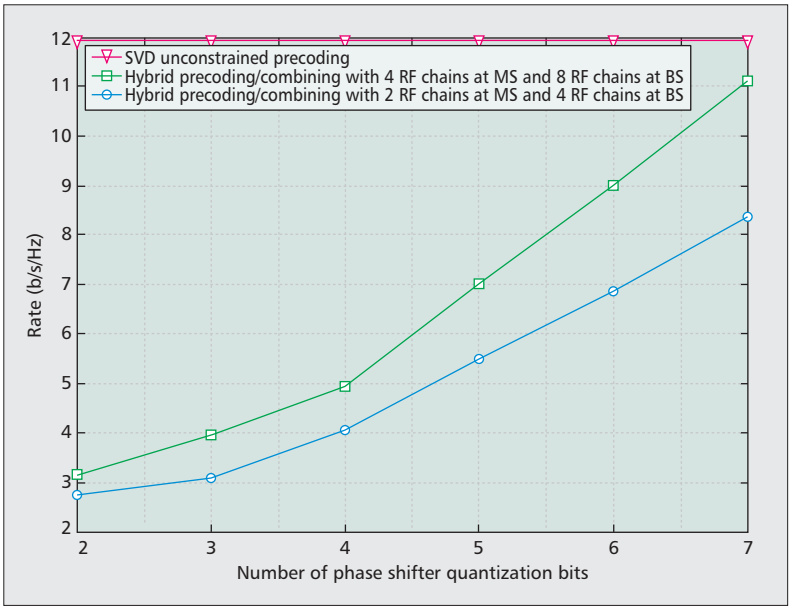


Figure 6. Spectral efficiency achieved by hybrid precoding with different numbers of quantization bits for the base station and mobile user analog phase shifters. In this figure, the hybrid precoding/combining of two setups of the base station and mobile user with different numbers of RF chains are illustrated and compared to the unconstrained SVD solution.

Future Research Directions

There are still many open problems that need to be investigated on precoding/combining architectures for mmWave communications. We highlight several potential directions below.

Multi-User Mmwave Systems with Hybrid Precoding

Currently, most of the research done on hybrid precoding design has focused on single-user systems [5, 6]. Extensions to multi-user mmWave systems is of great interest. Hybrid precoding enables different beams to be assigned to different users through the analog precoding layer, and allows for more processing to be done in the digital layer to manage the interference between users [15]. It is therefore interesting to employ hybrid precoding in multi-user mmWave systems. Developing multi-user hybrid precoding and combining matrices, though, is challenging given the different hardware constraints. Considering out-of-cell interference in the design of hybrid precoding schemes for mmWave cellular systems is interesting. Finding efficient ways to divide the required multi-user precoding and scheduling processing between the analog and digital domains will be very useful to enable mmWave cellular systems.

Compressed Sensing Estimation of Multi-User Mmwave Channels

In [6], the sparse nature of the mmWave channel has been exploited to realize efficient estimation algorithms that require low training overhead. This training overhead, however, scales linearly with the number of users. Hence, developing efficient multi-user channel estimation algorithms is essential. Random beamforming transmission and compressed sensing tools provide a promising research direction to design mmWave channel estimation algorithms that allow all users to simultaneously estimate the channel, and hence decrease the associated training overhead. Research is still needed, however, to develop specific algorithms and study their performance.

Performance Analysis for Combining with > 1-Bit ADCs

The 1-bit approach has the advantage of extremely low power consumption. At the same time, the achievable rate is limited by the extremely coarse quantization. With 2-bit ADCs at the receiver, the maximum achievable rate will double from $2N_r$ b/s/Hz to $4N_r$ b/s/Hz. Hence, there is interest in extending the analysis

mmWave precoding is still an active area of research. Continued effort is needed to develop generalizations of hybrid and 1-bit ADC precoding to multi-user mmWave systems and to design efficient training and channel estimation algorithms that incur low overhead in the presence of mobility.

on 1-bit quantization to the case of ADCs with resolution larger than 1 bit (e.g., 2–3). For the case of 2–3 bits of quantization, one of the main difficulties is to find the optimal quantization thresholds. First, the uniform quantization may not be optimal. Second, the optimal thresholds will change with the dynamic range of the received signal. Third, for the MIMO channel, optimizing the thresholds of each ADC separately seems challenging. A possible simplification is to assume that each ADC has the same thresholds.

Training Signal Design for Systems with 1-Bit ADCs

In contrast with unquantized systems, the capacity of a communication link with quantization is achieved by a discrete input distribution due to the nature of discrete quantization outputs [7]. Therefore, instead of designing the transmitted signal to estimate the exact channel state, it may be of interest to only estimate the channel responses when certain discrete symbols are transmitted. Based on the possible optimal inputs, typical discrete symbols are chosen as the training signals. For example, in the low SNR region, independent quadrature phase shift keying (QPSK) signaling across different antennas has near-optimal performance. Therefore, in this case, the training signals can be chosen to be QPSK symbols. With these estimated channel responses, the receiver can detect which QPSK symbol is transmitted on each of the transmitting antennas. The complexity and performance of this approach need further exploration. At medium and high SNR, QPSK signaling is expected to have worse performance than the optimal approach, and choosing the best training signals remains an open problem.

Precoding and Combining Strategies for the Broadband Mmwave Channel

In previous sections we have discussed different precoding/combining techniques designed and tested under the assumption of a narrowband channel model. Further research is needed to consider the broadband scenario, which implies different statistical models for the channel parameters: angles of arrival, path losses, or multipath time delay spreads requiring equalization.

Alternative Receiver Architectures

In this article we have reviewed the design and performance of three different architectures: analog-only beamforming, hybrid precoding and combining, and combining with low-resolution ADCs. Other design strategies that also make use of the sparse nature of the mmWave channel are possible as well. For example, receiver architectures based on the idea of randomly switching antennas instead of using phase shifters at the analog combining layer could be designed. In this way, the analog combiner could be seen as a random measurement matrix, and compressive sensing theory could be applied to estimate the channel.

Conclusions

Precoding and receiver design will be an important component of future mmWave cellular systems, and the next generation of mmWave wireless local area network standards. We present two precoding/combining strategies that take the different hardware constraints, different antenna scales, and different channel characteristics into consideration, making them suitable for operation in mmWave systems. Performance examples show that these solutions compare favorably with unconstrained optimal precoding strategies despite the reduction in cost and complexity. mmWave precoding is still an active area of research. Continued effort is needed to develop generalizations of hybrid and 1-bit ADC precoding to multi-user mmWave systems and to design efficient training and channel estimation algorithms that incur low overhead in the presence of mobility.

Acknowledgment

This material is based on work supported in part by the National Science Foundation under Grant Nos. 1218338 and 1319556, and by a gift from Huawei Technologies, Inc.

References

[1] T. Rappaport *et al.*, *Millimeter Wave Wireless Communications*, Prentice Hall, 2014.

[2] Z. Pi and F. Khan, "An Introduction to Millimeter-Wave Mobile Broadband Systems," *IEEE Commun. Mag.*, vol. 49, no. 6, 2011, pp. 101–07.

[3] T. Bai, A. Alkhateeb, and R. Heath, "Coverage and Capacity of Millimeter-Wave Cellular Networks," *IEEE Commun. Mag.*, vol. 52, no. 9, Sept. 2014, pp. 70–77.

[4] T. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, 2013, pp. 335–49.

[5] O. El Ayach *et al.*, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, Mar. 2014, pp. 1499–1513.

[6] A. Alkhateeb *et al.*, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, Oct. 2014, pp. 831–46.

[7] J. Singh, O. Dabeer, and U. Madhow, "On the Limits of Communication with Low-Precision Analog-to-Digital Conversion at the Receiver," *IEEE Trans. Commun.*, vol. 57, no. 12, 2009, pp. 3629–39.

[8] M. R. Akdeniz *et al.*, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," arXiv preprint arXiv:1312.4921, 2013.

[9] J. Wang *et al.*, "Beam Codebook Based Beamforming Protocol for Multi-Gbps Millimeter-Wave WPAN Systems," *IEEE JSAC*, vol. 27, no. 8, 2009, pp. 1390–99.

[10] X. Zhang, A. Molisch, and S. Kung, "Variable-Phase-Shift-Based Rf-Baseband Codesign For MIMO Antenna Selection," *IEEE Trans. Signal Processing*, vol. 53, no. 11, 2005, pp. 4091–4103

[11] B. Le *et al.*, "Analog-to-Digital Converters," *IEEE Signal Processing Mag.*, vol. 22, no. 6, 2005, pp. 69–77.

[12] J. Mo and R. Heath, "High SNR Capacity of Millimeter Wave MIMO Systems with One-Bit Quantization," *Proc. Info. Theory and Applications Wksp.*, 2014.

[13] A. Mezghani, F. Antreich, and J. Nossek, "Multiple Parameter Estimation with Quantized Channel Output," *Proc. 2010 Int'l. ITG Workshop on Smart Antennas*, 2010, pp. 143–50.

[14] S. Suh *et al.*, "Low-Power Discrete Fourier Transform for OFDM: A Programmable Analog Approach," vol. 58, no. 2, 2011, pp. 290–98.

[15] A. Alkhateeb, G. Leus, and R. W. Heath Jr., "Limited Feedback Hybrid Precoding for Multi-User Millimeter Wave Systems," submitted to *IEEE Trans. Wireless Commun.*, arXiv preprint arXiv:1409.5162, 2014.

Biographies

Ahmed Alkhateeb [S'10] (aalkhateeb@utexas.edu) received his B.S. (with highest honors) and M.S. degrees from Cairo University, Egypt, in 2008 and 2012, respectively. He is currently a Ph.D. student in the Wireless Networking and Communication Group (WNCG), University of Texas at Austin. His research interests are in the broad area of network information theory, communication theory, and signal processing. In the context of wireless communication, his interests include cooperative communications, MIMO systems, and mmWave communication.

Jianhua Mo [S'12] (jhmo@utexas.edu) received his B.S. and M.S. degrees in electronic engineering from Shanghai Jiao Tong University in 2010 and 2013, respectively. He also received an M.S. degree in electrical and computer engineering from Georgia Institute of Technology. He is currently a Ph.D. student in WNCG, University of Texas at Austin. His research interests include physical layer security and millimeter wave communications.

Nuria González-Prelcic (nuria@gts.uvigo.es) is currently an associate professor in the Signal Theory and Communications Department, University of Vigo, Spain. She got her Ph.D. degree in telecommunications engineering from the University of Vigo in 1998, distinguished with the best Ph.D. thesis award. She has held visiting positions with Rice University (1997), the University of New Mexico (2012), and the University of Texas at Austin (2014). Her main research focus is on signal processing for communications. She is currently the head of the Atlantic Research Center for Information and Communication Technologies (AtlantTIC), and coordinator of the Research Cluster "Technological Progress and Competitiveness," both at the University of Vigo.

Robert W. Heath, Jr. [Fi11] (rheath@utexas.edu) is a Cullen Trust Endowed Professor in the Electrical and Communications Engineering Department at the University of Texas, Austin, and director of WNCG. He received B.S. and M.S. degrees in electrical engineering from the University of Virginia, and his Ph.D. in electrical engineering from Stanford University. He is the president and CEO of MIMO Wireless Inc. and chief innovation officer at Kuma Signals LLC. He is a registered Professional Engineer in Texas. He is co-author of the textbook *Millimeter Wave Wireless Communications* (Prentice Hall, 2014).

IEEE 802.11ad: Directional 60 GHz Communication for Multi-Gigabit-per-Second Wi-Fi

Thomas Nitsche, Carlos Cordeiro, Adriana B. Flores, Edward W. Knightly, Eldad Perahia, and Joerg C. Widmer

INVITED PAPER

ABSTRACT

With the ratification of the IEEE 802.11ad amendment to the 802.11 standard in December 2012, a major step has been taken to bring consumer wireless communication to the millimeter wave band. However, multi-gigabit-per-second throughput and small interference footprint come at the price of adverse signal propagation characteristics, and require a fundamental rethinking of Wi-Fi communication principles. This article describes the design assumptions taken into consideration for the IEEE 802.11ad standard and the novel techniques defined to overcome the challenges of mm-Wave communication. In particular, we study the transition from omnidirectional to highly directional communication and its impact on the design of IEEE 802.11ad.

INTRODUCTION

With the worldwide availability of a large swath of spectrum at the 60 GHz band for unlicensed use, we are starting to see an emergence of new technologies enabling Wi-Fi communication in this frequency band. However, signal propagation at the 60 GHz band significantly differs from that at the 2.4 and 5 GHz bands. Therefore, efficient use of this vast spectrum resource requires a fundamental rethinking of the operation of Wi-Fi and a transition from omnidirectional to directional wireless medium usage. The IEEE 802.11ad amendment addresses these challenges, bringing multi-gigabit-per-second throughput and new application scenarios to Wi-Fi users. These new uses include instant wireless synchronization, high-speed media file exchange between mobile devices without fixed network infrastructure, and wireless cable replacement (e.g., to connect to high definition wireless displays).

The most significant difference in 60 GHz propagation behavior is increased signal attenuation. At a typical IEEE 802.11ad range of 10 m, additional attenuation of 22 dB compared to the 5 GHz band is predicted by the Friis transmission equation, resulting from the frequency-dependent difference in antenna aperture. In contrast, oxygen absorption plays a minor role over short-range distances, even though it peaks at 60 GHz [1]. Furthermore, 60 GHz communication is characterized by a quasi-optical propagation behavior [2] where the received signal is dominated by the line of sight (LOS) path and first order reflections from strong reflecting materials. As an example, metallic surfaces were found to be strong reflectors and allow non-LOS (NLOS) communication [2]. Concrete materials, on the other hand, cause additional large signal attenuation and can easily create a blockage. Thus, 60 GHz communication is more suitable to in-room environments where sufficient reflectors are present.

This article discusses the design assumption resulting from the millimeter-wave (mm-Wave) propagation characteristics and related adaptation to the 802.11 architecture. We further present typical device configurations, an overview of the IEEE 802.11ad physical (PHY) layer, and the newly introduced personal basic service set network architecture. This is followed by an in-depth description of the IEEE 802.11ad beamforming (BF) mechanism and hybrid medium access control (MAC) design, which are the central elements to facilitate directional communication.

DIRECTIONAL COMMUNICATION

The IEEE 802.11ad amendment to the 802.11 standard defines a directional communication scheme that takes advantage of beamforming antenna gain to cope with increased attenuation in the 60 GHz band [1]. With quasi-optical propagation behavior, low reflectivity, and high attenuation, beamforming results in a highly directional signal focus. Based on this behavior, the standard introduces a novel concept of "virtual" antenna sectors [3] that discretize the antenna azimuth. IEEE 802.11ad sectors can be implemented either using precomputed antenna weight vectors for a phased antenna array [4] or equipping a system with multiple directional antenna elements. In both cases, the wavelength in the millimeter range allows antenna form fac-

Thomas Nitsche is with IMDEA Networks Institute and University Carloss III of Madrid.

Joerg C. Widmer is with IMDEA Networks Institute.

Carlos Cordeiro and Eldad Perahia are with Intel Corporation.

Adriana B. Flores and Edward W. Knightly are with Rice University.

0163-6804/14/$25.00 © 2014 IEEE

Device	Antenna sectors	Expected range (m)	Expected maximum throughput (Gb/s)	Traffic type	Antenna arrays
AP, docking station	32 to 64	20	7	Bursty traffic on downlink	≤3
Wireless peripheral (hard drive, memory stick)	≤ 4	0.5 to 2	4.6	Bursty	1
Wireless display, TV	32 to 64	5 to 10	7	Continuous, RX more important	≤2
Notebook	16 to 32	5 to 10	4.6 to 7	Various, symmetric TX and RX	≤2
Tablets	2 to 16	2 to 5	4.6	Various, symmetric TX and RX	1
Smartphone, handheld, camcorder, camera	≤ 4	0.5 to 2	1.2 to 4.6	Various, symmetric TX and RX, TX more important for video streaming devices	1

Table 1. Typical device configurations.

tors significantly smaller than those of legacy Wi-Fi at 2.4/5 GHz.

A sector focuses antenna gain in a certain direction. Communicating nodes thus have to agree on the optimal pair of receive and transmit sectors to optimize signal quality and throughput. This process, referred to as beamforming training, takes advantage of the discretized antenna azimuth that reduces the search space of possible antenna array configurations. After a first sector matching, a second beam training stage allows further refinement of the found sectors. During this stage, antenna weight vectors that vary from predefined sector patterns can be evaluated to further optimize transmissions on phased antenna arrays. While in general higher antenna gain is desirable, it imposes stronger directionality and a higher number of narrow antenna sectors. This increases coordination overhead to adapt the antenna steering between communicating nodes, and it has been shown that link budget loss by misalignment increases with directionality [5].

Figure 1 shows an example for two nodes communicating over virtual IEEE 802.11ad sectors. The highlighted selection of sectors that matches the LOS direction may offer the optimum link quality in the absence of blocking obstacles.

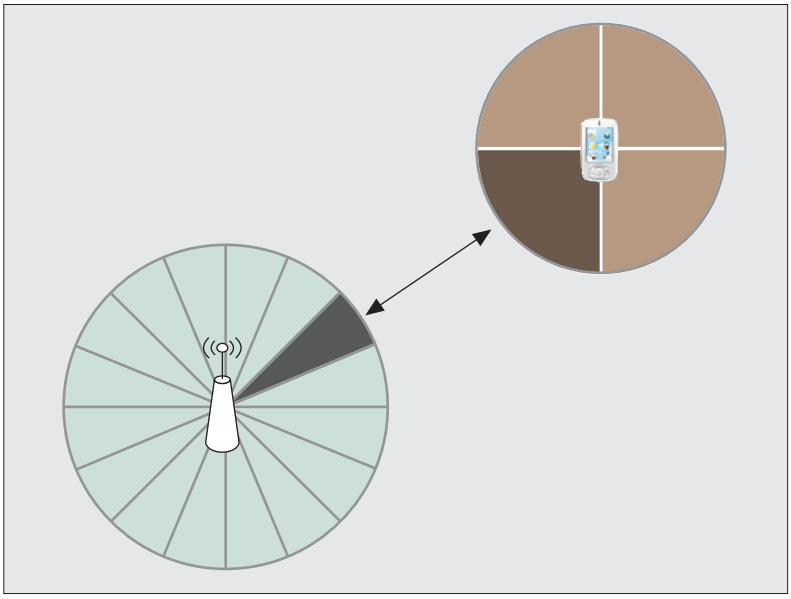


Figure 1. Virtual antenna sectors.

IEEE 802.11AD Device Classes and Use Cases

Communication in the mm-Wave band enables extremely high throughput at short-ranges (< 10 m), with high potential for spatial reuse. Thus, not only does it suit typical Wi-Fi usage, it also expands the uses of Wi-Fi to other application areas. Among these areas are wireless transmissions of high definition video, wireless docking stations, connection to wireless peripherals, or high-speed download of large media files.

To meet the requirements for these novel use cases, the IEEE 802.11ad standard allows for a broad variety of directional multi-gigabit (DMG) devices ranging from energy constrained handheld equipment with low complexity antennas (1–4 antenna elements) to stationary access points with multiple antenna arrays and permanent power supply. Table 1 shows typical configurations for several device classes. It states the number of sectors that correlates with range and throughput, differences between receive and transmit direction, and special traffic characteristics for every class. Furthermore, the expected

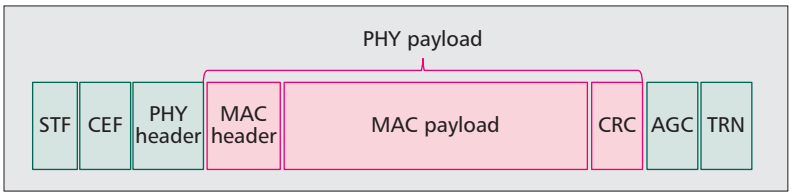


Figure 2. IEEE 802.11ad packet structure.

number of antenna arrays is given for every device class. Multiple phased antenna arrays enable high gain coverage in all directions. They are not used in a multiple-input multiple-output (MIMO) fashion, but treated like a set of additional sectors with only one antenna array used at a time.

IEEE 802.11AD Design Assumptions

Communication in the mm-Wave frequency band has different characteristics than legacy 2.4/5 GHz Wi-Fi frequencies. Thus, the development of the IEEE 802.11ad amendment followed a number of design assumptions that result from the change of frequency band.

Highly directional transmissions: Increased transmission loss and the application of high gain beamforming techniques lead to a strong directional signal focus. In contrast to omnidirectional legacy Wi-Fi signal propagation, IEEE 802.11ad communicates over narrow beams that follow quasi-optical propagation characteristics.

Quasi-omnidirectional antenna patterns: Implementation of truly omnidirectional mm-Wave antenna patterns is not practical, as signal blockage and deviation by device components in the vicinity of the antenna have a much stronger effect than on legacy Wi-Fi frequencies. Therefore, IEEE 802.11ad introduces quasi-omnidirectional patterns that allow gain fluctuations over the pattern. Further measures are taken to cope with the resulting inaccuracies.

Inefficient omnidirectional communication: The increased attenuation in the mm-Wave band leads to severely reduced transmission range and throughput when quasi-omnidirectional antenna patterns are used. However, when the direction to a communication partner is unknown (e.g., during beamforming training), quasi-omni patterns are still needed. Thus, directional antenna gain is added at least at one side of a link to achieve a sufficient communication range. Typically, quasi-omnidirectional antenna configurations are used at the receiver side. Only devices with extreme space or energy restrictions are expected to implement quasi-omnidirectional transmit modes. These devices will be severely limited in range and throughput (Table 1).

Extreme efficiency loss on poorly trained beams: The throughput difference between the highest and lowest transmission scheme defined by IEEE 802.11ad lies in the range of 6.5 Gb/s. A poorly trained beam that uses a low-throughput scheme severely reduces the system performance and should be avoided at all costs.

Reduced interference footprint: Highly directional transmission properties of IEEE 802.11ad devices strongly reduce interference outside of the beam direction. This allows spatial reuse of the same frequency band and can significantly increase the system's overall throughput.

Deafness and directional communication drawbacks: Highly directional IEEE 802.11ad transmissions have hindering effects on common Wi-Fi MAC mechanisms. Directional transmit patterns prevent devices from passively overhearing ongoing transmissions, leading to additional collisions during channel access. Furthermore, the deafness effect caused by misaligned transmit or receive antenna patterns may lead to frame loss, unnecessary long contention backoff, and lower throughput. An in-depth discussion of these impairments can be found in [6]. IEEE 802.11ad adapts the 802.11 carrier sense multiple access with collision avoidance (CSMA/CA) mechanism and further introduces a multi-MAC architecture, with alternative medium access schemes suited to directional communication.

IEEE 802.11AD Physical Layer

IEEE 802.11ad introduces three different PHY layers dedicated to different application scenarios. The *control PHY* is designed for low signalt-to-noise ratio (SNR) operation prior to beamforming. The *single carrier (SC)* PHY enables power-efficient and low-complexity transceiver implementation. The low-power SC PHY option replaces the low-density parity check (LDPC) encoder by a Reed-Solomon encoder for further processing power reduction. The orthogonal frequency-division multiplexing (*OFDM*) *PHY* provides high performance in frequency selective channels achieving the maximum 802.11ad data rates.

Despite having different PHYs, all of them share the same packet structure with common preamble properties. Specifically, the same Golay sequences are used for the preamble training fields. Also, a common rate 3/4 LDPC structure is used for channel encoding. Moreover, 802.11ad defines a single bandwidth of 2.16 GHz, which is 50 times wider than the channels available in 802.11n and roughly 14 times wider than the channels defined in 802.11ac.

The single IEEE 802.11ad packet structure is shown in Fig. 2. The packet consists of typical IEEE 802.11 elements, for example, a short training field (STF) and a channel estimation field (CEF) that is also used for auto-detection of the PHY type. They are followed by the PHY header and PHY payload, which is protected by a cyclic redundancy check (CRC). Finally, optional automatic gain control (AGC) and training (TRN) fields, unique to IEEE 802.11ad, might be appended. These are used for the beamforming mechanism described later.

To provide robust discovery and detection, the control PHY has a longer STF than the SC and OFDM PHYs, comprising 48 Golay sequences, each 128 samples long. The SC and OFDM PHY only use 17 Golay sequences for the STF. The channel estimation field that follows the STF has nine Golay sequences. The OFDM PHY uses a different combination of

Golay sequences in the CEF to distinguish between OFDM and SC modulation.

The control PHY defines modulation and coding scheme (MCS) 0. It implements a 32-sample Golay spreading sequence along with rate 1/2 LDPC encoding (spread from the common rate 3/4 LDPC code) to extend range and reliability for management frames, giving a throughput of 27.5 Mb/s. The control PHY uses π/2-differential binary phase shift keying (BPSK) modulation to further enhance robustness to distortion like phase noise. The mandatory control PHY defines the minimum rate all devices may use to communicate before establishing a high-rate beamformed link. It is used for transmitting and receiving frames such as beacons, information request and response, probe request and response, sector sweep, sector sweep feedback, and other management and control frames.

The SC PHY (MCS 1-12) and low-power SC PHY (MCS 25-31) allow for low-complexity and energy-efficient transceiver implementations with a throughput of up to 4.62 Gb/s. The lowest SC data rate is 385 Mb/s (MCS 1). It is implemented using BPSK modulation and rate 1/2 code with a symbol repetition of two. All modulation types use π/2 rotation to reduce the peak-to-average power ratio for BPSK and enable Gaussian minimum shift keying (GMSK) equivalent modulation. To provide interoperability between different device types, MCS 1–4 are mandatory for all devices. These four MCSs are all based on π/2-BPSK modulation. MCS 2, 3, and 4 use code rate 1/2, 5/8, and 3/4, respectively.

The OFDM PHY (MCS 13-24) is an optional mode for maximum throughput at the cost of a more complex and energy intensive transceiver structure. The OFDM PHY type utilizes 64-QAM and a rate 13/16 code to achieve the highest 802.11ad data rates of up to 6.75 Gb/s.

In order to keep transceiver complexity and energy consumption low, mobile and low-cost devices are likely to implement only SC PHYs. In contrast, stationary devices with fixed power supply and high throughput requirements (access points, wireless displays) implement the full spectrum of MCSs including complex OFDM transceivers.

IEEE 802.11AD Network Architecture

This section describes the changes to the IEEE 802.11 network architecture defined by IEEE 802.11ad. First, we describe the changes to the beacon interval (BI). Next, a novel network type called personal basic service set (PBSS) is introduced, followed by the description of the network and schedule announcement mechanisms.

Beacon Interval

IEEE 802.11 in lower frequency bands organizes the medium access through periodically recurring beacon intervals that are initiated by a single beacon frame transmitted omnidirectionally by the access point (AP) or coordinating station. The beacon announces the existence of a Wi-Fi network and carries further management data.

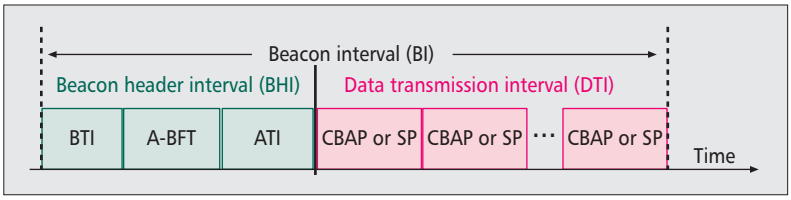


Figure 3. IEEE 802.11ad beacon interval structure.

The rest of the BI is used for data transmissions between stations, usually following a contention-based access scheme. The length of a BI is limited to 1000 ms, but typically chosen in the range of 100 ms. While longer BI durations increase the connection delay for nodes waiting for the beacon, a longer interval reduces management frame transmission and increases throughput.

The IEEE 802.11ad amendment to the IEEE 802.11 standard extends this concept in several ways to cope with the challenges of mm-Wave propagation. First, a BI is initiated with the *beacon header interval* (BHI), which replaces the single beacon frame of legacy Wi-Fi networks. The BHI facilitates the exchange of management information and network announcements using a sweep of multiple directionally transmitted frames. The BHI sweeping mechanism overcomes increased attenuation and unknown direction of unassociated devices. Additional functionality of the BHI is described later on. The BHI is followed by a *data transmission interval* (DTI), which can implement different types of medium access. The schedule and medium access parameters, which are necessary for stations to participate in a BI, are announced by the central network coordinator, the PBSS control point (PCP) or AP, during the BHI. This ensures that stations receive this information even though no efficient broadcasting mechanism is available.

A typical BI, consisting of BHI and DTI, is shown in Fig. 3. The BHI consists of up to three sub-intervals. First, the *beacon transmission interval* (BTI) comprises multiple beacon frames, each transmitted by the PCP/AP on a different sector to cover all possible directions. This interval is used for network announcement and beamforming training of the PCP/AP's antenna sectors. Second, the *association beamforming training* (A-BFT) is used by stations to train their antenna sector for communication with the PCP/AP. Third, during the *announcement transmission interval* (ATI), the PCP/AP exchanges management information with associated and beam-trained stations. While communication during BTI and A-BFT uses MCS 0 to increase range for untrained beams, communication during the ATI takes place with beam-trained stations and thus is more efficient.

The DTI comprises one or more *contention-based access periods* (CBAPs) and scheduled *service periods* (SPs) where stations exchange data frames. While in CBAP multiple stations can contend for the channel according to the IEEE 802.11 enhanced distributed coordination function (EDCF), an SP is assigned for communication between a dedicated pair of nodes as a contention-free period.

Dynamic channel time allocation is an extension of the IEEE 802.11 PCF mode. It provides higher flexibility in resource allocation (polled stations request channel time instead of just transmitting one frame) and adaptation to directional communication.

Personal Basic Service Set

IEEE 802.11ad introduces the PBSS, where nodes communicate in an ad hoc like manner. However, one of the participating nodes takes the role of the PCP. This PCP acts like an AP, announcing the network and organizing medium access. This centralized approach allows the directional network and schedule announcement process described in the next section to be used for an ad hoc like network. The PBSS network has been introduced to satisfy new applications targeted by IEEE 802.11ad, such as wireless storage and peripherals or wireless display usage. For these applications, usually no preinstalled infrastructure exists, and communication takes place between a set of personal devices.

An ad hoc like network with a centralized controller poses two main challenges. First, for energy-constrained devices, increased power consumption at the PCP penalizes a single device while fair sharing of the energy costs is desirable. Second, outage of the PCP paralyzes the entire PBSS. To respond to these challenges, a PCP handover procedure is defined [3]. This procedure can be used for explicit (initiated by the current PCP) or implicit (after a PCP becomes unavailable) handovers. Furthermore, when selecting between a set of possible PCPs, the unique capabilities of PCP candidate stations are considered to choose the PCP providing the most complete number of services to the network.

Network and Schedule Announcements

Network announcements in legacy IEEE 802.11 are traditionally propagated periodically, using beacon frames, by the AP. Due to the limited antenna gain of quasi-omnidirectional mm-Wave transmissions, the coverage range is severely restricted. Consequently, the beacon is sent as a series of directionally transmitted beacon frames. To have the largest possible range, the beacon frames are transmitted at the most robust MCS (MCS 0). IEEE 802.11ad also specifies additional signaling for network scheduling and beam training appended to every beacon frame. Collectively, this results in a significantly increased overhead in comparison to legacy Wi-Fi. Thus, it becomes critical to control the amount of information transmitted in each BTI. In addition, transmissions during the A-BFT, which also use MCS 0, create overhead recurring with every BI where the A-BFT is present. The overhead problem gets especially relevant when short BI durations are applied for delay-critical application such as video streaming.

The IEEE 802.11ad amendment defines a number of counter strategies. First, it is possible to split a beacon sweep over several BIs. This, however, increases the time a node needs to set up its link to the PCP/AP, as not every direction is served at every BI. The result is an increased association delay. Second, it is possible to periodically schedule BIs without A-BFT, which also results in additional association delays. Third, IEEE 802.11ad introduces the ATI. During the ATI, beam-trained and associated nodes can be served with management data using individually addressed directional transmitted frames encoded with a more efficient MCS. Thus, it is possible to move information from the spectrally inefficient beacon frames to the frames transmitted during the ATI, limiting beacons to the minimal information necessary.

Also, for beacon intervals with split beacon sweeps, stations that do not receive a beacon miss network and timing information. Without this information, stations cannot participate in a BI. Implementing an ATI solves this problem, as scheduling and management information is transmitted individually to associated stations.

IEEE 802.11AD MAC Layer

In contrast to legacy Wi-Fi, IEEE 802.11ad uses a hybrid MAC approach to address its various use cases [3, 7]. The standard supports contention-based access, scheduled channel time allocation, and dynamic channel time allocation. The latter two schemes correspond to time-division multiple access (TDMA) and polling mechanisms. Polling-based access shares similarities with the IEEE 802.11 point coordination function (PCF) mode, but is adapted to directional transmissions and provides a higher flexibility when it comes to distribution of resources among the nodes. The scheduled allocation mechanism extends the traffic stream concept known from the IEEE 802.11 hybrid coordination function (HCF) to request time shares of the DTI for TDMA-like medium access. Next, the three methods are described.

Contention-Based Medium Access

Medium access in CBAPs follows IEEE 802.11 enhanced distributed channel access (EDCA), including traffic categories to support quality of service, frame aggregation, and block acknowledgments. However, when using contention-based access with directional antennas, the problem of deafness arises. A deaf node does not receive directionally transmitted information due to misaligned antenna patterns. A detailed description of the effect can be found in [6]. While the beam training process in IEEE 802.11ad prevents deafness for intended transmissions, it poses a problem for carrier sensing during contention-based access and can lead to increased collisions. A further problem for contention-based access is that a receiver typically does not know where a signal comes from. Thus, usage of quasi-omnidirectional beam patterns is necessary, which reduces link budget and throughput.

The contention-based medium access in IEEE 802.11ad is adapted for directional medium usage and multi-MAC usage. This includes support for multiple network allocation vector (NAV) timers (one per peer station), which allows a transmission to be initiated to a peer device where the NAV for that device is zero, even though the NAV for another peer device might be nonzero. Details about 802.11 EDCA and its use in 802.11ad can be found in [3, 8].

Dynamic Channel Time Allocation

IEEE 802.11ad defines a dynamic channel time allocation mechanism that implements polling-based channel access. Dynamic channel time allocation is an extension of the IEEE 802.11 PCF mode. It provides higher flexibility in resource allocation (polled stations request

channel time instead of just transmitting one frame) and adaptation to directional communication. Polling-based channel access brings several advantages for mm-Wave communication. First, due to the centralized approach with a PCP/AP, stations are aware of the direction of incoming signals. Thus, the deafness problem that affects contention-based access is prevented, and quasi-omnidirectional receive patterns can be avoided. Second, centralized scheduling at a PCP/AP also helps to efficiently react to bursty downstream traffic, as dynamic scheduling can be adapted in the course of a BI. In contrast, pseudo-static scheduling, described in the following section, can only announce modified allocation parameters with the beginning of every BI.

When applying the dynamic allocation mechanism, the medium access during DTI is organized as follows. The PCP/AP acquires the medium and sends a series of polling frames to associated stations. This is answered with a block of service period requests (SPRs) used by the polled stations to request channel time. The PCP/AP allocates the available channel time according to these requests, announcing each allocation with a separate grant period, consisting of individual grant frames for the stations involved in the allocation.

IEEE 802.11ad foresees integration of the dynamic allocation mechanism into both CBAPs and SPs. When integrated in a CBAP, associated stations try to acquire the medium and may interfere with the dynamic allocations. To prevent this, the PCP/AP makes use of prioritized medium access using the short PIFS inter-frame spacing, and the channel is protected by extension of the frame duration fields. This extension causes nodes that overhear a frame to assume that the channel is occupied until the time specified in the duration field. This mechanism is used such that polling and SPR frames protect the polling phase, while every dynamic allocation is protected by its preceding grant frames.

To simplify the scheduling mechanism and reduce implementation complexity, dynamic allocations are scheduled back to back, with every allocation immediately following its Grant period. To reliably reach the nodes that are involved in an allocation, individual directional frames are transmitted during the grant period. In case of an allocation between PCP/AP and a station, only one grant frame is sent to the non-PCP/AP station.

When not all available channel time is allocated dynamically, the PCP/AP can repeat the entire polling process. For integration into a CBAP, remaining channel time can also be used for CSMA/CA access.

An example for three polled stations is shown in Fig. 4. The PCP/AP commences a polling phase at the beginning of the DTI, transmitting a polling frame for every associated station, which is answered with a series of three SPRs by the stations. The second station requests communication with another non-PCP/AP station, while stations one and three intend to communicate with the AP (not shown). The resulting allocations are scheduled back to back, each preceded by a grant period. For communication with the AP, the grant period consists of one frame; otherwise, two. The time until which preceding frames protect the channel is indicated by separating lines.

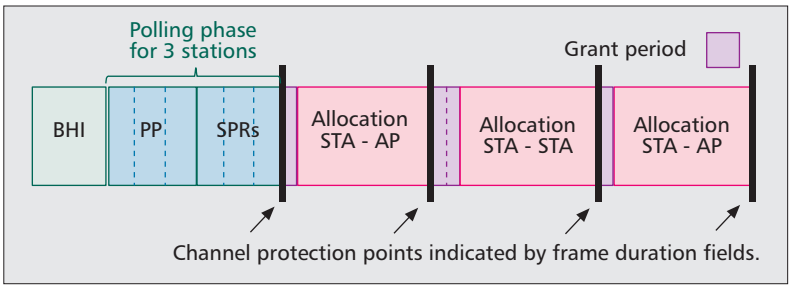


Figure 4. Dynamic channel allocation.

Pseudo-Static TDMA Channel Time Allocation

During pseudo-static channel time allocation, SPs that recur every BI are dedicated exclusively to a pair of communicating nodes. Accessing the channel using this TDMA mechanism provides reliability and is the best way to comply with quality of service demands. Furthermore, the schedule of SPs is propagated by the PCP/AP to all associated stations. Thus, every node that is not communicating during a SP can go into sleep mode, which allows efficient power saving.

For pseudo-static medium allocation, the concept of traffic streams for IEEE 802.11 HCF, as described in [8], is extended. A traffic stream is defined as a flow of MAC service data units that has to be delivered subject to certain quality-of-service parameters characterized by a traffic specification.

The IEEE 802.11ad amendment defines stations to use traffic specifications to request scheduling of pseudo-static channel allocations at the PCP/AP. A requesting station defines the properties of its traffic demand in terms of allocation duration and isochronous or asynchronous traffic characteristic. Calculating the allocation duration requires a complete beam-trained link with known rate between source and destination. Otherwise, the traffic specification has to be modified after beam training when the link's throughput rate is known. An isochronous traffic stream results in pseudo static SP allocations that satisfy a constant rate of recurring payload (typical, e.g., for wireless display applications) with certain latency demands. Asynchronous traffic streams, in contrast, satisfy non-recurring payload demand. A typical example application is rapid file download.

The actual schedule that includes the requested allocations is broadcasted by the PCP/AP in an extended schedule element in the next BTI or ATI.

IEEE 802.11ad Beamforming Concept

Beamforming training determines the appropriate receive and transmit antenna sectors for a pair of stations. This is achieved by transmission of a bidirectional training frame sequence. Throughout the training process, double-sided

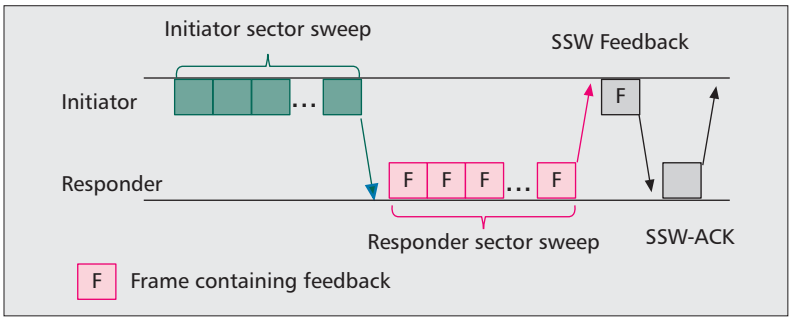


Figure 5. Sector-level sweep.

omnidirectional transmissions are avoided as they are severely limited in range.

The beamforming phase is split into two subphases. First, during the sector-level sweep (SLS), an initial coarse-grained antenna sector configuration is determined. This information is used in a subsequent optional beam refinement phase (BRP), which fine-tunes the selected sectors. During SLS each of the two stations trains either its transmit antenna sector or the receive antenna sector. When devices are capable of reasonable transmit antenna gain, the most common choice is to train only transmit sectors during SLS and derive receive antenna configuration during a following BRP. Fully refined transmit and receive sectors at both sides of a link allow multi-gigabit-per-second speeds to be reached over ranges up to 10 m.

This section explains the general approach to beamforming introduced in the IEEE 802.11ad standard. The beamforming concept allows a significant amount of implementation-dependent customization and has a variety of optional features. Therefore, we first focus on the mandatory SLS phase followed by a description of the mandatory parts of the BRP.

Sector-Level Sweep Phase

During the SLS, a pair of stations exchanges a series of sector sweep (SSW) frames (or beacons for transmit sector training at the PCP/AP) over different antenna sectors to find the one providing the highest signal quality. During the SLS, each station acts once as a transmitter and once as a receiver of a sweep, as shown in Fig. 5. The station that transmits first is called the initiator, the second the responder. Both initiator and responder sweep can be used in two different ways, as depicted in Fig. 6. During a transmit sector sweep (TXSS), shown in the left part of the figure, frames are transmitted on different sectors while the pairing node receives with a quasi-omnidirectional pattern. To identify the strongest transmit sector, the transmitter marks every frame with an identifier for the used antenna and sector. During a receive sector sweep (RXSS), shown in the right part of Fig. 6, transmission on the same sector (best known sector) allows to test for the optimum receive sector at the pairing node. Overall, there are four possible sweep combinations for an SLS. Transmit sector sweeps at both initiator and responder receive sector sweeps at both stations, initiator RXSS and responder TXSS, and initiator TXSS and responder RXSS.

The achieved optimum SNR and, in the case of a TXSS, the sector and antenna identifier are reported to the pairing node. SLS feedback follows the structure described in Fig. 5.

Feedback for the initiator is carried by every frame of the responder sector sweep, which ensures reception under still unknown optimum antenna configuration. The feedback for the responder is transmitted with a single SSW Feedback frame on the determined optimum antenna configuration. Finally, the SSW Feedback frame is acknowledged with an SSW-ACK by the responder. The last frame is further used to negotiate the details of a following BRP.

If two stations have sufficient transmit antenna gain, their SLS phase can be realized as pure transmit sector training, with the receive sector training postponed to a following BRP. Devices with few antenna elements have to add antenna gain at the receiver side in order to achieve sufficient link budget to establish a link. Thus, these devices are likely to include a receive sector sweep in their part of the SLS.

The initiator can request that a receive sector sweep be done by the responder by specifying the number of receive sectors to train during the initiator sweep. When the initiator sweep is a receive sector training, additional signaling has to precede the SLS, as described later.

The Beam Refinement Protocol Phase

The BRP refines the sectors found in the SLS phase. These sectors are determined using heterogeneous quasi-omnidirectional antenna patterns and may have suboptimal signal quality. Furthermore, the BRP foresees optimization of antenna weight vectors, independent of the predefined sector patterns, for phased antenna arrays. This can yield additional throughput gains while increasing the beam training search space. Even though free variation of the antenna weight vectors can result in arbitrary antenna patterns, the directional nature remains for antenna configurations that yield high throughput. Thus, the training process for predefined directional sectors and antenna weight vector optimization remains the same. Finally, the BRP is used to train receive antenna configurations in case this was not part of the preceding SLS. Multiple optional pattern refinement mechanisms are defined for the BRP and are out of scope of this article. We focus on the mandatory beam refinement transactions, an iterative process in which both initiator and responder can request training for receive or transmit antenna patterns.

A BRP transaction evaluates a set of directional transmit or receive patterns against the best known directional configuration at the pairing node. Thus, the imperfection of quasi-omnidirectional patterns is avoided. As the BRP relies on a preceding SLS phase, a reliable frame exchange is ensured, and different antenna configurations can be tested throughout the same frame. This severely reduces transmission overhead in contrast to the SLS, where a full frame is necessary to test a sector. To sweep antenna configurations throughout a frame, transmit and receive training fields (TRN-T/R) are appended to the frames exchanged during BRP transactions. Each field is transmitted or received with

an antenna configuration that is to be tested for its signal quality. The remaining portion of the frame is transmitted and received with the best known antenna configuration.

BRP receive antenna training is requested by specifying the number of configurations to be tested in a frame's L-RX header field. The pairing node will append the according number of TRN-R fields to its next frame. A transmit training is requested by setting the TX-TRN-REQ header field and appending TRN-T fields to the same BRP frame. Optionally, no training fields are attached, and an acknowledgment frame with the TX-TRN-OK field set is transmitted by the recipient before the requester appends the TRN-T fields to its following frame. Equal to the SLS, BRP feedback is given in form of SNR for the best found configuration and the best configuration ID in case offor a transmit training.

Figure 7 shows a BRP transaction that first trains the receive configuration between two stations, followed by additional transmit training refinement. Note that station B combines the request for transmit and receive training in one frame using the request variation explained above. Station A, in contrast, uses two frames to request the two transmit directions. The frames and training fields belonging to one of the different training requests are marked in the same color.

A BRP phase can immediately follow the SLS, using the SSW ACK frame for parameter exchange. Alternatively, it can be initiated by a special BRP setup sub-phase, consisting of training-field-free BRP frames. In either case, L-RX and TX-TRN-REQ fields are used to exchange the BRP parameters.

IEEE 802.11AD Beamforming Protocol

The general beamforming concept described earlier integrates into the different IEEE 802.11ad medium access schemes and the association processes. Before association, stations use an adapted version of the beamforming process to connect with the PCP/AP without preceding coordination. This training is further realized in a way that allows the PCP/AP to do sector training to all stations at the same time rather than separately.

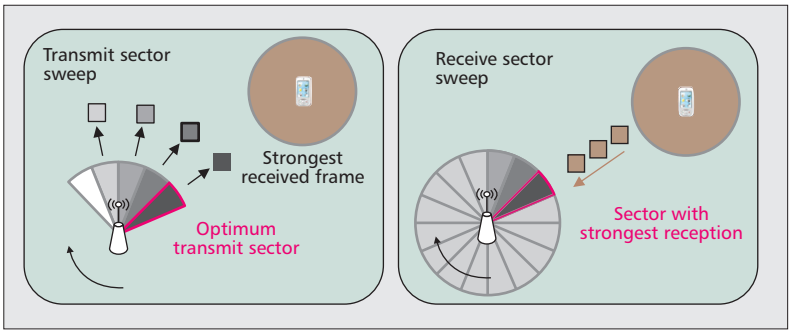


Figure 6. Transmit and receive sector training.

This section explains the association beamforming training, followed by a description of beam training between non-PCP/AP stations in accordance with the three different MAC schemes.

Association Beamforming Training

Beamforming training between the PCP/AP and an unassociated station cannot rely on coordination preceding to the beam training. To overcome the challenges of directional link setup, the PCP/AP uses its beacon sweep during the BTI, as an initiator sector sweep for all stations. To this aim, SSW frame specific control fields are added to the beacon frame. To allow multiple stations to respond to a beacon sweep without coordination, the A-BFT interval implements a contention-based response period. The A-BFT reserves channel time for multiple responder sector sweeps (A-BFT slots) from the stations. An overview for the association beamforming training during BTI and A-BFT is shown in the upper left of Fig. 8.

Each A-BFT slot consists of a fixed time allocation for a number of SSW frames (transmitted by the connecting station) and one SSW Feedback frame sent by the PCP/AP as depicted in the lower part of Fig. 8. Contending stations randomly select which slot to access.

The contention process during an A-BFT does not apply carrier sensing. Instead, a colli-

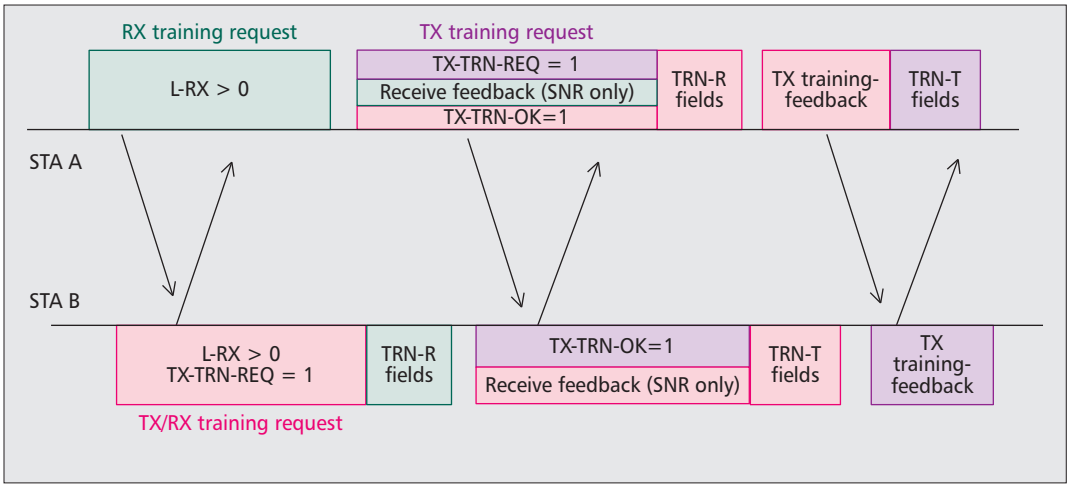


Figure 7. Beam refinement transactions.

Beamforming training during the DTI can be initialized following two different methods. First, the initiator can directly begin a sector level sweep when it gains control over the channel. Second, the PCP/AP can convey beam training parameters between two nodes, during dynamic or pseudo-static channel allocation.

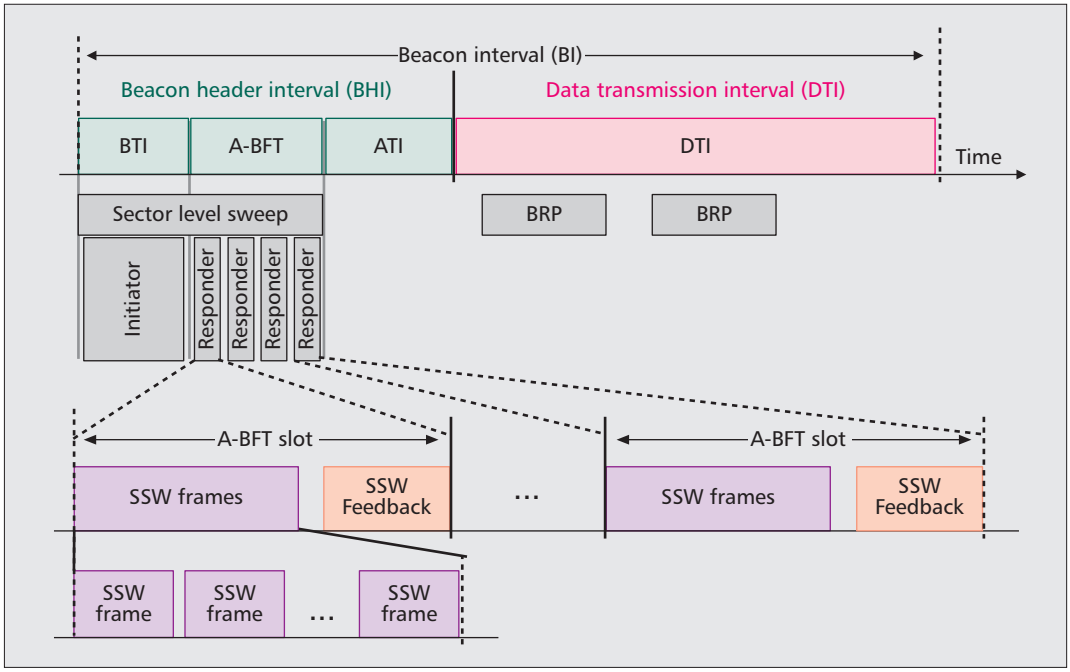


Figure 8. Association beamforming training.

sion is detected by a missing SSW Feedback frame from the PCP/AP. In addition, a station might be unable to finish its sweep because its sectors exceed the number of SSW frames per slot. To handle such cases, several measures can be taken. First, the PCP/AP can answer an incomplete sweep with an SSW Feedback frame, forcing the selection of a suboptimal transmit sector. Second, a station might contend for further slots during the A-BFT in the same or a following BI. To resolve congestion of the association beamforming training interval, a station has to draw an additional amount of backoff slots when its retries exceed a given limit. Also, the beam training can be moved into a dedicated SP by the PCP/AP according to the procedures described below. BRPs for the links between the PCP/AP and stations are scheduled in the DTI, as indicated in the upper right of Fig. 8.

A PCP/AP can announce an A-BFT for receive sector training. Hereby, the slot size indicates the number of receive sectors the PCP/AP intends to train, and associating stations transmit the according number of SSW frames.

Beam Training in the Data Transmission Interval

Beamforming training during the DTI can be initialized following one of two methods. First, the initiator can directly begin an SLS when it gains control over the channel. This method is required during CSMA/CA access. Second, the PCP/AP can convey beam training parameters between two nodes, during dynamic or pseudo-static channel allocation. Using the second mechanism, the PCP/AP learns about the pending beam training and can integrate that information into the scheduling process.

For direct beam training initialization, a station that has seized the channel initiates the beamforming process with a transmit sector sweep to the responder. However, if the initiator intends to start receive antenna training, additional signaling is necessary. In that case, the initiator inquires the number of receive sectors at the responder via the PCP/AP or higher-level protocols. Then, to initialize the SLS, a Grant/Grant-ACK exchange is used to request a receive sector sweep. Following that, both nodes start the training after the Grant-ACK frame. During contention-based access, short inter frame spacing between beamforming frames ensures that no other node wins a transmit opportunity and causes interference.

Beam training via the PCP/AP during pseudo static channel allocation is requested with the initial traffic specification that is transmitted. The beam training parameters are included by the PCP/AP in the extended schedule element that announces the first allocation, which causes both nodes of a traffic stream to commence training at the beginning of their first allocation.

To initiate beam training via the PCP/AP during dynamic channel allocation, a node requests an allocation to the beam training partner. In its corresponding SPR frame, the initiator indicates the parameters for the intended training. When granting the corresponding allocation request, the PCP/AP includes the beam training parameters into the Grant frames sent to both stations involved in the allocation.

Beam refinement during the DTI typically follows immediately after a SLS. The initiator uses the SSW ACK frame to request transmit or receive training as described above. A station that has seized the channel can also initiate a standalone BRP using a BRP setup phase. To request mandatory beam refinement transactions only, the setup phase comprises a single BRP frame initiating the refinement sequence.

Conclusions

In this article, we present the IEEE 802.11ad standard, which brings consumer wireless communication to the millimeter wave band. We highlight the standard's hybrid MAC layer design that defines three different medium access schemes: CSMA/CA, Polling, and TDMA. Every scheme addresses different aspects of mm-Wave communication and supports varying quality of service mechanisms, making it suitable for different IEEE 802.11ad use cases.

Furthermore, we address the elaborate beam training protocol, which enables highly directional communication. The association beamforming training and two-level beam training are the fundamental elements of this protocol. First, association beam training aligns antenna beams between a station and a central network controller while the direction between the two devices is unknown. Second, two-level training reduces the beam training search space using its primary coarse-grained training stage that relies on predetermined virtual antenna sectors. Its second stage further refines the found antenna configuration varying from predefined sectors and also addresses the challenges of imperfect omnidirectional antenna patterns. With fully trained transmit and receive antenna configurations, IEEE 802.11ad reaches its maximum throughput of up to 7 Gb/s. In addition, the beamforming protocol supports a training procedure for low antenna gain devices and can convey training parameters to a central network coordinator for channel access scheduling.

The combination of the hybrid MAC layer and the novel beam training protocol is key to enabling new IEEE 802.11ad use cases, and addressing specific device and millimeter wave propagation characteristics.

References

[1] P. Smulders, "Exploiting the 60 GHz Band for Local Wireless Multimedia Access: Prospects and Future Directions," *IEEE Commun. Mag*., vol. 40, no. 1, Jan. 2002, pp. 140–47.

[2] H. Xu, V. Kukshya, and T. Rappaport, "Spatial and Temporal Characteristics of 60-GHz Indoor Channels," *IEEE JSAC*, vol. 20, no. 3, April, 2002, pp. 620–30.

[3] IEEE 802.11 WG, "IEEE 802.11ad, Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band," Dec. 2012.

[4] A. Valdes-Garcia *et al.*, "Single-Element and Phased-Array Transceiver Chipsets for 60-GHz Gb/s Communications," *IEEE Commun. Mag*., vol. 49, no. 4, Apr. 2010, pp. 120–31.

[5] H. Yang, P. Smulders, and M. Herben, "Frequency Selectivity of 60-GHz LOS and NLOS Indoor Radio Channel," *Proc. IEEE VTC*, May 2006.

[6] R. Choudhury and N.H. Vaidya, "Deafness: A MAC Problem in Ad Hoc Networks when using Directional Antennas," *Proc. ICNP*, Oct. 2004.

[7] C. Cordeiro, "Evaluation of Medium Access Technologies for Next Generation Millimeter-Wave WLAN and WPAN," *Proc. IEEE ICC Wksps.*, June, 2009.

[8] IEEE 802.11 WG, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," Mar. 2012.

Biographies

THOMAS NITSCHE (thomas.nitsche@imdea.org) is a Ph.D. student with IMDEA Networks Institute and Universidad Carlos III, Madrid, Spain. In 2009 he received a Diploma degree in computer sciences from Technische Universität München, Germany, and was a Ph.D. candidate at the Chair for Network Architectures and Services until he joined IMDEA Networks Institute in 2012. His research focuses on design and implementation of wireless PHY and MAC layer protocols, mm-Wave Wi-Fi, wireless localization systems, and cross-layer protocol design.

CARLOS CORDEIRO [SM] (carlos.cordeiro@intel.com) is a principal engineer in the Mobile and Communications Group within Intel Corporation. He leads Intel's standardization programs in Wi-Fi and in the area of short-range multi-Gb/s wireless systems using millimeter frequencies. In the Wi-Fi Alliance, he is a member of the Wi-Fi Alliance Board of Directors and serves as the Wi-Fi Alliance Technical Advisor, in addition to chairing the technical task group on 60 GHz. He was the technical editor of the IEEE 802.11ad standard. Due to his contributions to wireless communications, he has received several awards including the prestigious Global Telecom Business 40 under 40 in 2012 and 2013, the IEEE Outstanding Engineer Award in 2011, and the IEEE New Face of Engineering Award in 2007. He is the co-author of two textbooks on wireless published in March 2006 and March 2011, has published about 100 papers in the wireless area alone, and holds over 30 patents. He has served as Editor of various journals.

ADRIANA B. FLORES (a.flores@rice.edu) is a Ph.D. candidate in the Department of Electrical and Computer Engineering at Rice University. She received her M.S. in electrical engineering from Rice University in 2012 and her B.S.E.E in 2009 from Monterrey Institute of Technology and Higher Education (ITESM), Mexico. She joined the Rice Networks Group in 2011, where she works under the guidance of Dr. Edward Knightly, and holds a Texas Instruments Distinguished Fellowship. Her research focuses on the design of medium access control protocols for efficient channel usage in both sub-GHz and mm-Wave band.

EDWARD W. KNIGHTLY [F] (knightly@rice.edu) is a professor of electrical and computer engineering at Rice University. He received his Ph.D. and M.S. from the University of California at Berkeley, and his B.S. from Auburn University. He is a Sloan Fellow and a recipient of the NSF CAREER Award. His group's current projects include deployment, operation, and management of a large-scale urban wireless network in an underresourced community in Houston, Texas. The network is the first to provide residential access in frequencies spanning from unused UHF DTV bands to Wi-Fi bands, and employs custom-built programmable and observable access points.

ELDAD PERAHIA (eldad.perahia@intel.com) is a principal engineer in the Mobile and Communications Group within Intel Corporation and is the physical layer lead in IEEE 802.11. He is currently engaged in the new high-efficiency WLANs (802.11ax) activity and regulatory affairs. He was the 802.11ad (60 GHz) Chair and 802.11aj Vice-Chair (mmWave in China). He was also actively involved in the 802.11ac task group since its inception. He was the 802.11ac Coexistence Ad Hoc Co-Chair. Prior to that, he was Chair of the 802.11 Very High Throughput Study Group that launched 802.11ac and 802.11ad. He was also the 802.11n Coexistence Ad Hoc Chair. He is the co-author of *Next Generation Wireless LANs: 802.11n and 802.11ac* (Cambridge, 2013). He has 25 patents, and numerous papers and patent filings in various areas of wireless including WLAN, millimeter wave technology, satellite communications, cellular, and radar. He has a Ph.D. from the University of California, Los Angeles in electrical engineering specializing in digital radio.

JOERG WIDMER [SM] (joerg.widmer@imdea.org) is a research professor at IMDEA Networks Institute, Madrid, Spain. He received his M.S. and Ph.D. degrees in computer science from the University of Mannheim, Germany, in 2000 and 2003, respectively. His research focuses primarily on wireless networks, ranging from MAC layer design and interference management to mobile network architectures. From 2005 to 2010, he was manager of the Ubiquitous Networking Research Group at DOCOMO Euro-Labs, Munich, Germany, leading several projects in the area of mobile and cellular networks. Before, he worked as a post-doctoral researcher at EPFL, Switzerland, on ultra-wide band communication and network coding. He was a visiting researcher at the International Computer Science Institute in Berkeley, California, and University College London, United Kingdom. He has authored more than 100 conference and journal papers and three IETF RFCs, holds several patents, serves on the Editorial Board of *IEEE Transactions on Communications*, and regularly participates in program committees of several major conferences. Recently, he was awarded an ERC consolidator grant as well as a Spanish Ramon y Cajal grant. He is a Senior Member of ACM.

The combination of the hybrid MAC layer and the novel beam training protocol is key to enabling new IEEE 802.11ad use cases and addressing specific device and millimeter wave propagation characteristics.

TRENDS IN CONSUMER COMMUNICATIONS



Ali C. Begen *Mario Kolberg* *Madjib Merabti*

The last 15 years have heralded many developments and advances in consumer communications, from early developments of device-specific challenges in interoperability and configuration that are well captured by the concept of plug and play to a more recent emphasis on mobility and service personalization. The one constant technical challenge, and to a great extent a business success, is home networking in its many forms. There is not a modern home without some variant of a set-top box. The three articles in this issue provide a good overview of current and topical requirements in consumer communications.

The first article, "Homenet3D: A New View on Home Network State" by Armitage *et al.*, reviews the state of the art in home networking from a very apt point of view: the consumer and system's usability. The challenge of managing and operating more complex home networks is discussed in detail before a new approach using visualization techniques is proposed.

The second article, "Greening the Spectrum: A Minority Game Based Mechanism Design" by Elmachkour *et al.*, addresses the dual challenge of what to do with the crowding of the communication spectrum and energy consumption. In this article the authors propose techniques of cognitive radio and minority game-based mechanisms in order to reuse underutilized frequency spectrum and reduce energy consumption.

The third article, by Nightingale *et al.*, "Video Adaptation for Consumer Devices: Opportunities and Challenges Offered by New Standards," discusses the latest developments in video compression technology standards that are aimed at improving service quality. In particular, it focuses on the use of new adaptation techniques.

If the articles in this series are of interest to you, we strongly urge you to consider participating in the IEEE Consumer Communications and Networking Conference (CCNC) 2015 that will be held next January in Las Vegas in conjunction with the Consumer Electronics Show (CES), the largest CE show in the world. See http://www.ieee-ccnc.org for details.

BIOGRAPHIES

ALI C. BEGEN [SM] (abegen@cisco.com) is with the Video and Content Platforms Research and Advanced Development Group at Cisco. His interests include networked entertainment, Internet multimedia, transport protocols, and content distribution. He is currently working on architectures for next-generation video transport and distribution over IP networks, and he is an active contributor in the IETF in these areas. He holds a Ph.D. degree in electrical and computer engineering from Georgia Tech. He received the Best Student Paper Award at IEEE ICIP 2003 and the Most Cited Paper Award from Elsevier *Signal Processing: Image Communication* in 2008. Recently, he was General Co-Chair for the ACM Multimedia Systems Conference 2011. He organized a special session on IPTV and related technologies at Packet Video Workshop 2012. Further information on his projects, publications, and presentations can be found at http://ali.begen.net.

MARIO KOLBERG [SM] (mkolberg@ieee.org) is a senior lecturer within the Institute of Computing Science and Mathematics at the University of Stirling, United Kingdom. His research interests include peer-to-peer overlay networks, home automation, and IP telephony. He is on the Editorial Board of the Springer journal *Peer-to-Peer Networking and Applications* and has a long standing involvement with IEEE CCNC. He served as TPC Chair of the January 2011 conference. Currently, he is chairing the track on Human Centric Computing at IEEE GLOBECOM 2014. He has published more than 50 papers in leading journals and conferences. He is a member of a number of international ' program committees on networking and communications. He holds a Ph.D. from the University of Strathclyde, United Kingdom.

MADJID MERABTI [M] (M.Merabti@ljmu.ac.uk) is a professor of networked systems and director of the School of Computing and Mathematical Sciences at Liverpool John Moores University, United Kingdom. He is currently on leave as Dean of the College of Sciences at the University of Sharjah, United Arab Emirates. He holds a Ph.D. from Lancaster University, United Kingdom. He has over 20 years' experience in conducting research and teaching in the areas of computer networks (fixed and wireless), mobile computing, and computer network security. He is widely published, with over 150 publications in these areas, and leads the Distributed Multimedia Systems and Security Research Group. He is principal investigator in a number of current projects: Mobile Networks Security and Privacy Architectures and Protocols, Secure Component Composition in Ubiquitous Personal Networks, Networked Appliances, Mobile and Ad Hoc Computing Environments, Sensor Networks, and computer games technology. He was Guest Editor for a Special Issue on Research Developments in Consumer Communications and Networking of *Multimedia Tools and Applications: An International Journal* (Kluwer, September 2005). He is a member of the Steering Committee for IEEE CCNC. He has acted as TPC chair for a number of international conferences, including the 5th IEEE Workshop on Networked Appliances, Liverpool, United Kingdom, October 2002. He is a member of a number of international conferences program committees on networking, security, and computer entertainment.

Homenet3D: A New View on Home Network State

Grenville Armitage and Dominic Allan

ABSTRACT

Although it performs an increasingly important role in our lives, the home network remains a bit of a mystery when end users wish to know what is "happening on the inside." This article introduces Homenet3D, an open source project that allows users to view their home network's state as objects in 3D space within a web browser window. Homenet3D maps quantitative state to the shape, size, spin, bounce, and/or color of selected objects to qualitatively communicate what is happening on the network. We describe our Homenet3D implementation for routers running OpenWRT, and discuss the potential for visualizing multi-router/multi-subnet home networks.

INTRODUCTION

Today's modern, Internet-connected residential home has an increasingly busy network within its borders. Typically there will be a fixed broadband connection into the home, providing Internet access to PCs, laptops, tablets, smartphones, network-enabled TVs, streaming media clients, security cameras, and so on. As the *Internet of Things* (IoT) takes hold, everyday devices (e.g., light switches, washing machines, air conditioning systems) are also being augmented with IP-based network interfaces and remote control systems [1]. However, despite its increasingly important role in our lives, the home network remains a bit of a mystery when end users wish to know what is "happening on the inside."

Our premise is that a useful, qualitative, relatively non-technical view of the home network's current state may be created through the use of suitably designed and animated objects in a virtual 3D environment. Furthermore, network configuration changes might be effected through an end user's interaction with objects inside the virtual world.

The Homenet3D project [2] envisages the typical home user utilizing her/his laptop, tablet, or smartphone to view their home network's state as objects in 3D space within a browser window. Particular visual attributes of these objects (e.g., shape, size, spin, and/or color) will vary over time to qualitatively communicate what is happening on the network.

The rest of this article is structured as follows. The following section places Homenet3D within the broader context of efforts to define and manage home networks. The next section discusses the state of 3D rendering support in consumer devices. Our prototype of Homenet3D for OpenWRT is introduced after that, and configuration options are then described. We go on to explore the implications of multi-router home networks, controlling network devices through the 3D interface, and potential applications outside the home environment. Our conclusions are presented in the final section.

HOME NETWORKS

For many current homes, a single IP gateway sits between the home and the broadband modem providing asymmetric digital subscriber line (ADSL), data over cable system interface spectrum (DOCSIS), optical fiber, or fixed wireless service to the outside world. Communication within the home will involve one or more link layer technologies such as wired Ethernet, 802.11 WiFi, or Homeplug (power line) networking. The gateway, wired switch ports, 802.11 access point (AP), and Homeplug interfaces may be separate devices or integrated within the home's broadband modem.

Despite the diverse layer 2 technologies, current consumer products often encourage a single-subnet topology centered around a single home gateway. Network address translation (NAT) then enables all the in-home devices to (relatively transparently) communicate with the wider Internet using the home's single Internet service provider (ISP)-provided IP address.

Industry efforts, such as the Home Gateway Initiative (http://www.homegateway.org) and the UPnP Forum (http://www.upnp.org), are developing technologies for home networks to self-configure and operate in as much of a plug-and-play manner as possible. Not surprisingly, existing standards largely assume IPv4-based home networks. More recently the Internet Engineering Task Force's (IETF's) Homenet working group (http://tools.ietf.org/wg/homenet) has explicitly chosen to focus on an IPv6-based future where homes may have multiple internal subnetworks and routers, and support multiple ISP connections to the outside world (multihoming) [3].

The authors are with Swinburne University of Technology.

 0163-6804/14/$25.00

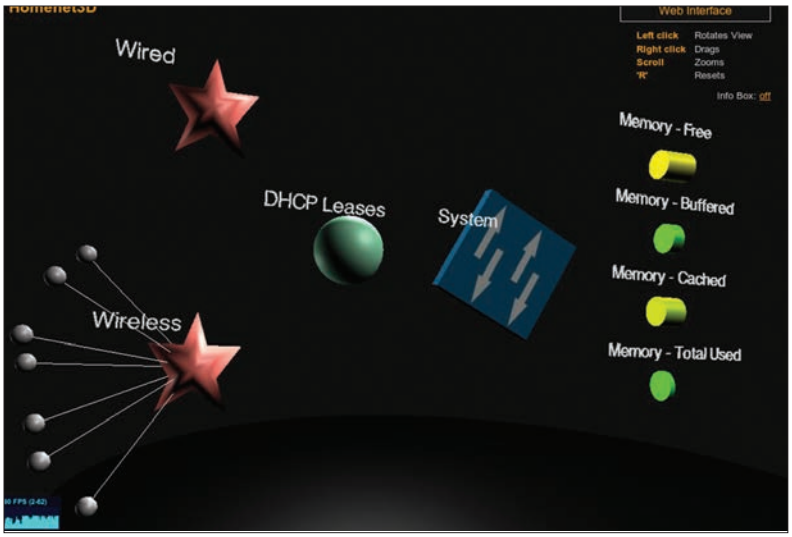


Figure 1. Homenet3D for OpenWRT viewed on a WebGL-enabled browser. Hovering the mouse over an object reveals detailed state information.

Homenet3D is not directly related to these other efforts. Rather than developing protocols for instantiating home networks, Homenet3D focuses on providing non-technical users with a qualitative sense of their home network's internal state. This might include the aggregate bandwidth in or out of the home, clients currently associated with their WiFi AP, Dynamic Host Configuration Protocol (DHCP)-assigned internal IPv4 or IPv6 addresses, ad hoc or auto-configured internal IPv6 hosts, active NAT sessions, firewall rules, and so on. Homenet3D augments current browser-based interfaces to home gateways, which usually provide terse and limited quantitative details (largely cryptic to the average non-technical end user).

Immersive 3D Monitoring

Dynamic 3D environments enhance the information that we may concurrently present to a viewer. Relationships may be qualitatively expressed by the spatial positioning of visible 3D objects relative to each other (distant vs. close, clustered vs. scattered, etc.). Underlying quantitative values may be approximated by a 3D object's static characteristics (scale, radius, color, or shape) or dynamic characteristics (e.g., rotation speed or bounce rate).

Presenting system state in 3D is not an intrinsically new idea. The SGI File System Navigator turned up in 1993's *Jurassic Park*, sporting a navigatable 3D representation of the UNIX file system. Subsequent examples include network activity visualization [4], virtual world metaphors for interacting with computer process space [5], and virtual world collaboration systems [6]. A recent survey of many ideas in 2D and 3D visualization to assist with network security monitoring can be found in [7].

From Game Engines to WebGL

Some of our own previous work focused on repurposing the server and client engines of a late-1990s era first person shooter (FPS) game to provide a multi-party monitoring of network state using 3D "worlds" [8]. Inside the virtual world, multiple network states were represented in real time using 3D objects with *visually orthogonal* attributes. (An example of visual orthogonality might include making a pyramid shaped object's spin rate and color proportional to packets per second and number of flows at a measurement point, respectively, as spin is unlikely to be mistaken for color.) We utilized the Quake III Arena game engine for efficient use of network resources between clients and server, and 3D-capable clients across multiple platforms [9].

However, dedicated multiplayer game clients are no longer the only place to find 3D rendering and networking capabilities. The emergence of HTML5 and WebGL standards opens up the potential for 3D virtual environments being created in-browser on modern consumer devices. Javascript toolkits like Three.js (http://threejs.org) ease the creation of dynamic and interactive in-browser 3D applications. HTML5's WebSockets (http://tools.ietf.org/html/rfc6455) enable browser-based applications to engage in continuous two-way communication with remote servers. In other words, new technologies are making it easier for residential home gateways to present their dynamic state inside 3D worlds displayed by HTML5/WebGL-compliant browsers.

The Rise of Smart Phones and Tablet Computers

Ten years ago we still needed decent PCs or laptops to provide high-resolution color graphics and significant computing power. Today, sub-$150 smartphones and tablet computers are sporting color touch screens with 800 × 600 or higher resolutions, hardware accelerated graphics, 802.11g/n WiFi connectivity, and 1 GHz+ processors.

As a productivity enhancer, fashion statement, and the focal point of many people's social lives, we might reasonably assume one or more such devices will be present in many home networks. Given the steadily increasing availability of mobile browsers capable of WebGL-based 3D rendering, Homenet3D arguably moves from theoretically interesting to plausibly useful.

Homenet3D for OpenWRT

To demonstrate Homenet3D we targeted OpenWRT (https://openwrt.org), a Linux distribution commonly used to replace the factory firmware on many commercial residential WiFi gateways and similar embedded systems. Replacing the factory firmware typically provides users with additional configuration options and extended functionality.

We have Homenet3D running on ARM and MIPS machines emulated in QEMU, and a physical TP-Link WR-1043ND wireless gateway. The OpenWRT SDK can build Homenet3D-capable firmware for other platforms that OpenWRT supports.

System State as Objects in a 3D World

Figure 1 illustrates a Homenet3D client (web browser window) presenting the network as seen by a single OpenWRT router and WiFi

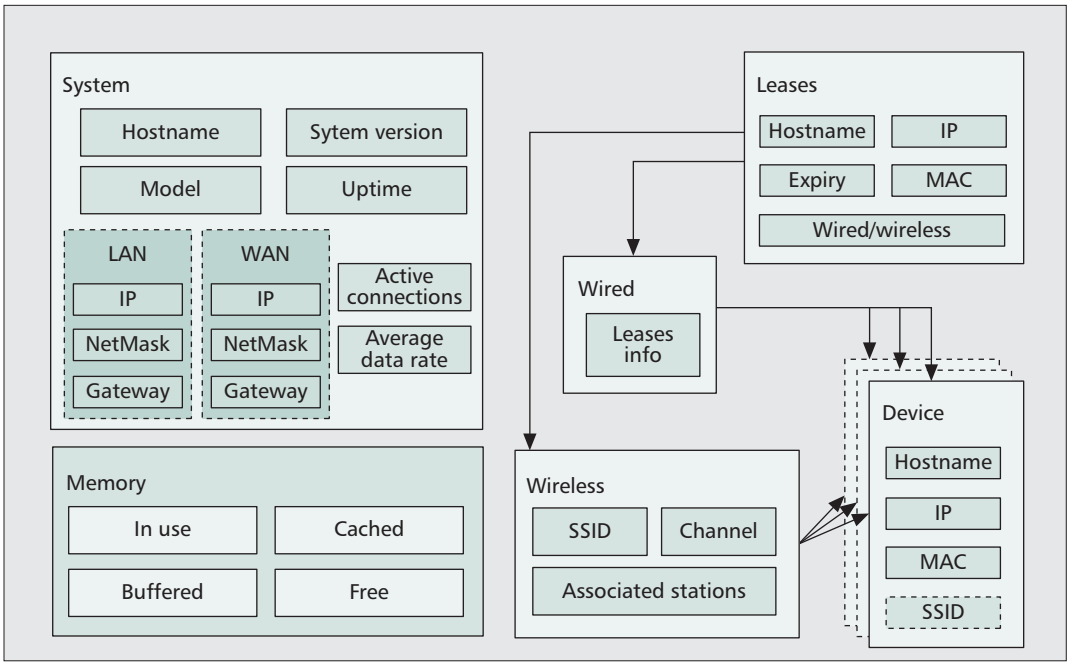


Figure 2. OpenWRT system states tracked for Homenet3D.

access point (AP). Different internal OpenWRT system states are mapped to individual on-screen *entities*. Homenet3D's entities are objects the appearance and behavior of which qualitatively represent some underlying system state. The user's mouse (or touch screen controls) are used to rotate, zoom, and pan the view. Hovering the cursor over an entity triggers a pop-up subwindow with more specific (quantitative) information about the associated system state.

Figure 2 shows the categories of OpenWRT system state tracked by Homenet3D. There is one entity each for system, memory, DHCP leases, aggregate wired devices, aggregate wireless devices, and individual wireless devices. Each entity is associated with specific information. For example, for the system entity there is the device's hostname and uptime as well as the network interface information for WAN and LAN. Wired and wireless entities house all information about the wireless and wired devices currently connected to the system. Both the wired and wireless entities spawn child device entities that detail each individual attached device. The wired, wireless, and device entities leverage the lease information housed in the leases entity.

Homenet3D currently provides a number of specific mappings. The blue switch model's spin rate and size is proportional to network traffic on the WAN interface. Each memory category is presented along a cylinder color-coded to represent different types of memory use. Wireless clients of the AP and wired DHCP clients are represented by small spheres on strings hanging off two large stars. More spheres cluster around each star as more clients arrive. Hovering one's mouse over either wired or wireless stars will bring up detailed information on client medium access control (MAC) addresses and assigned IP addresses in a pop-up window. Hovering one's mouse over an individual sphere on a string will bring up information about that specific client.

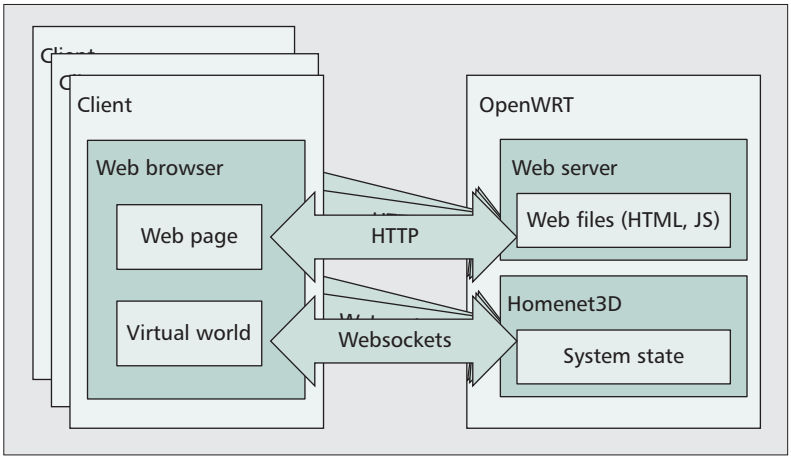


Figure 3. Multiple Homenet3D clients may concurrently view system state.

Turning Browsers into Homenet3D Clients

Viewing is as simple as pointing your browser at a specific page hosted by the OpenWRT device's internal web server (e.g., http://192.168.0.1/hnet3d if your OpenWRT gateway sits at 192.168.0.1). The returned index.html directs your browser to download and execute additional Javascript files from the OpenWRT device via regular HTTP on port 80. From this point on the browser is executing the Homenet3D client, and communication continues via WebSockets on port 10001.

As illustrated in Fig. 3, Homenet3D supports multiple Homenet3D clients being connected at the same time. Key OpenWRT system state is

Attribute	Format	Description
state_types	string vector	One or more internal states the entity represents
position	[x,y,z] vector	Object position in 3D space
bounce_height	integer	Height of object bounce in Z-axis
bounce_freq	integer	Period of bounce in Z-axis
rotate_speed	float	Rotation rate around Z axis
colour	[R,G,B] vector	Object color
label	string	Text displayed with object
text_colour	[R,G,B] vector	Text label color
radius	integer	Object radius
rotation	[rx,ry,rz] vector	Object's 3D orientation
shape	integer	Select predefined (1..6) or custom (7) shape
objfile, mtlfile	string	Names of custom OBJ and MTL files
scale	float	Ratio of original size to display size

Table 1. Configurable in-world object attributes.

streamed to all attached Homenet3D clients over their individual WebSocket connections. All 3D rendering and viewer controls are instantiated on the client side, allowing each end user to view and navigate around the Homenet3D world independently.

Resources Consumed by Homenet3D

OpenWRT is typically used on embedded systems having Flash memory in the 4–8 Mbytes range and 16 Mbytes or more RAM. In this context Homenet3D cannot afford to take up significant additional space. A monitoring system should also add little to existing home network traffic.

The current Homenet3D client consumes ~972 kbytes in the Flash image. A similar-sized burst of local network traffic is generated when downloading the Homenet3D client to a browser. (Many browsers will cache the initial download, eliminating much of this traffic from later reconnections to the Homenet3D server.) Subsequent traffic to each Homenet3D client is short ~5–10 kbyte bursts per change in monitored system state (e.g., when a WiFi client comes or goes). Such bursts of traffic will be fairly infrequent (tens of seconds or minutes between events).

Mapping State to 3D Representations

Homenet3D's ultimate utility depends on the meaningful mapping of network states (e.g., aggregate bandwidth or number of active WiFi clients) to the dynamic behavior of in-world entities representing those states.

Configuring Entities and Objects

A single configuration file on the Homenet3D server defines the mapping of different OpenWRT system states to on-screen entities, and the initial visual characteristics (attributes) of the 3D objects representing each entity (e.g., orientation, position, size, color, spin rate, and bounce rates). The configuration file further defines how changes in system state map to changes in one or more of an entity's attributes (e.g., how large an object will grow in response to increased bandwidth use). Table 1 shows a range of per-entity attributes that may be set.

We can illustrate this process using Algorithm 1, a fragment of Homenet3D configuration file defining the "switch" entity used to represent System state in Fig. 1. This entity is at location [0, 0, 0] in 3D space, rotated to a specific initial orientation [0.5, 0, 0] (expressed in radians around the x, y, and z axes, respectively) and has a text label of "System" written in 3D space next to the object. The `shape`, `objfile`, and `mtlfile` options indicate that the object itself is a custom design, and the `mappings=()` option indicates that the system's average data rate (Homenet3D's calculation of network traffic through the OpenWRT system's WAN interface) is mapped to the object's rotation speed.

Mapping States to Attributes

Homenet3D uses the `mappings=()` configuration option to define a list of linear mappings from measured system state values to displayed object attributes. Pairs of `metric={}` and `attrs=()` sub-parameters identify the system state and target attributes respectively.

For example, Algorithm 1 shows the System entity configured such that values of datarate (`metric_name`) between 300 (`lower_thresh`) and 100000 (`upper_thresh`) are mapped to a rotation speed (`attr_name`) between 1.0 (`attr_min`) and 5.0 (`attr_max`), with the rotation speed quantized to the nearest 0.5 (`attr_gran`). If the data rate falls outside either upper or lower thresholds, the object's rotation speed will be capped at the `max` or `min` rates, respectively.

Similar syntax can be used to establish other mappings between Homenet3D-monitored system state and numeric object attributes in Table 1. Being easily configurable, Homenet3D enables further research into the utility of different mapping strategies (e.g., whether spin, bounce, or color better capture the intuitive importance of state such as network traffic, memory consumption, or numbers of clients). The potential benefits of nonlinear mappings will also be explored in later versions of Homenet3D.

Using Custom 3D Objects

In addition to six predefined objects, Homenet3D allows the use of object geometry (OBJ) and material template (MTL) definition files to specify custom 3D objects for each entity. Created in the 1990s by Wavefront Technologies, OBJ and MTL files are easily generated by many 3D design tools such as Milkshape3D, Blender, Shade, and SketchUp Pro. (A diverse range of example objects can be found at sites such as http://tf3dm.com/3d-models/all.)

As illustrated in Algorithm 1, a custom object is selected using `shape=7`, and the associated OBJ and MTL files are identified using the `objfile` and `mtlfile` options. By default custom entity objects are stored on the OpenWRT device and retrieved during initial startup of each Homenet3D client.

Custom objects must be relatively simple and small to minimize consumption of the OpenWRT device's Flash memory. However, `objfile` and `mtlfile` may also be full URLs. This enables hosting of larger custom objects on another web server (also accessible from the home network) that has more local storage space and allows cross-origin content retrieval from Homenet3D clients.

```
name = "Homenet3D"
version = "0.3"
application =
{
// Global settings
text_colour = [211,211,211];
// Configuration example for system entity
 entities = (
 {
  position = [0,0,0];
  label = "System";
  state_types = ["system", "network"];
  shape = 7;
  objfile = "models/switch.obj";
  mtlfile = "models/switch.mtl";
  scale = 0.5;
  bounce_freq = 0.0;
  rotation = [0.5,0.0,0.0];
  // Mappings
  mappings = (
    {
      metric = {
        metric_name = "datarate";
        upper_thresh = "100000.0";
        lower_thresh = "300.0";
      }
      attrs = (
        {
          attr_name = "rotate_speed";
          attr_max = "5.0";
          attr_min = "1.0";
          attr_gran = "0.5";
        }
      )
    }
   );
  }
 );
}
```

Algorithm 1. Fragment of Homenet3D configuration file.

Visualizing Multi-Router Environments

Homenet3D can also encompass future homes that contain multiple internal networks, multiple routers, and (potentially) multiple connections to different ISPs.

Complex Home Environments

The single-gateway/single-subnet model is a convenient simplification that increasingly fails to reflect a modern home's network environment. It is increasingly likely that a home's primary network will be divided at the IP layer into wired and wireless subnetworks (e.g., to isolate guest and trusted WiFi access, or to separate home backup server traffic from IP telephony and games traffic). This implies multiple WiFi networks linking back to the common ISP connection, potentially multiple routers, and certainly more complex in-home topologies.

Complicating the situation further are networks created through *tethering*, where WiFi-equipped smartphones act as their own local WiFi access points and share their 3G or 4G cellular Internet access with other nearby WiFi devices. Such networks are transient and independent of the home's primary Internet access network. However, they can provide an entirely real service to the tethered WiFi devices (albeit slower, and with higher latencies, than Internet access through the home's fixed-line broadband service).

More advanced homes may also implement multihoming, where the primary in-house network has active gateways to multiple ISPs at the same time. Multihoming can ensure that a home network stays connected to the Internet when one (or more) ISPs suffer outages. The cost is additional complexity inside the home network: someone must make and distribute routing policies that control how the home's router(s) will spread traffic across the available ISP links.

More advanced homes may also implement multihoming, where the primary in-house network has active gateways to multiple ISPs at the same time. Multihoming can ensure that a home network stays connected to the Internet when one (or more) ISPs suffer outages.

Significant work is currently underway in the IETF's Homenet working group for automating the discovery and self-configuration of devices within multi-router, multi-subnet, and/or multihomed networks. Core to this is the Home Network Control Protocol (HNCP), both the specification [10] and an actual implementation for Linux routers (https://github.com/sbyx/hnetd) are under active development.

Presenting Multiple Routers with Homenet3D

Homenet3D provides one possible approach to visualizing the state of a network with multiple OpenWRT devices. First, we assume Homenet3D is installed on each OpenWRT device. In their Homenet3D configuration files, one OpenWRT device is designated the Homenet3D "master," while the other OpenWRT devices are designated Homenet3D "slaves." The master maintains system state information for itself and all slaves (who send the master regular updates of their own system states).

When a Homenet3D client connects to a Homenet3D master server, the user is presented with state information for all the slave devices rendered in 3D space. Figure 4 shows how this might look for a system of two slaves and one master router. Each Homenet3D master or slave

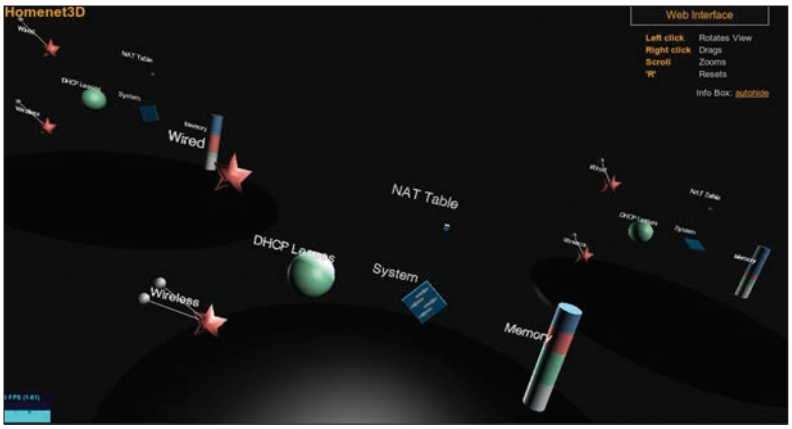


Figure 4. Homenet3D presenting a multi-router home network.

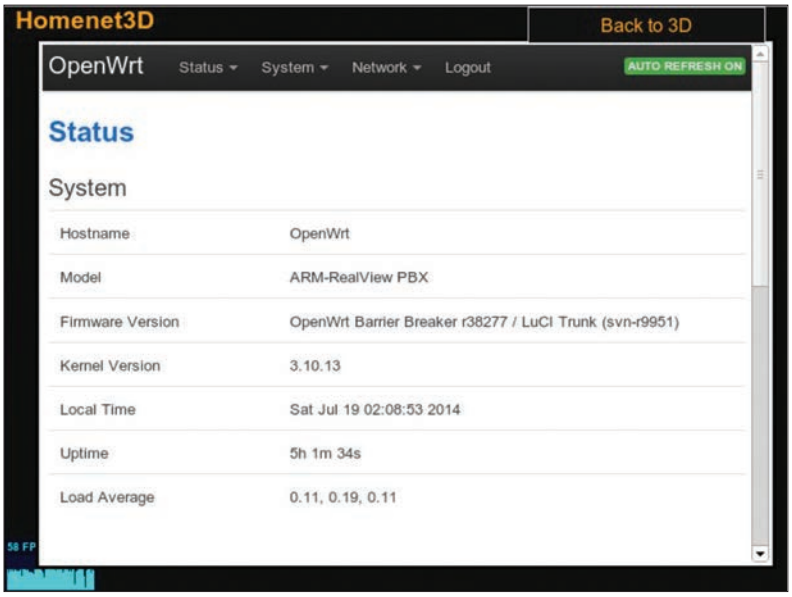


Figure 5. OpenWRT's conventional web interface inside Homenet3D.

has its own mappings between system state and on-screen entity attributes. Each device might, for example, use a different custom 3D object to represent their "system" state entity, or use different color, bounce, or spin rate mappings.

Each device must have Homenet3D preinstalled, with the identities of master and slave nodes specified in each devices' Homenet3D configuration file. As HNCP matures, future versions of Homenet3D will leverage HNCP functionality to auto-generate Homenet3D master and slave configuration files for discovered devices. As HNCP allows auto-discovery of network topology, we expect the spatial arrangement of entities within the Homenet3D world can reflect topological relationships between devices in the home network.

Interacting with Devices on the Home Network

The next obvious step is to offer *control* of network devices or systems through their in-world entities. We consider two avenues: control of actual home network infrastructure and control of household devices that are interconnected by the home network.

Home Network Infrastructure

Homenet3D currently provides easy access to OpenWRT's conventional web interface by clicking on the words Web Interface in the top right corner of the Homenet3D screen. The conventional interface is then presented on screen as shown in Fig. 5. After the user interacts in detail with the underlying OpenWRT router (potentially making specific changes to their home network), they click on Back to 3D to return to the 3D world view. The user may freely alternate between the 3D view and 2D GUI modes, with the conventional web interface remaining logged while the user is back in 3D view mode.

A key future enhancement will be to enable triggering of macros — pre-scripted control actions — when the user interacts with entities within the 3D view. For example, previous work using 3D game engines repurposed the in-game notion of "shooting" an object to trigger specific firewall rule updates, without the end user needing to worry about detailed firewall rule syntax [8].

In a Homenet3D environment we envisage pull-down menus appearing in 3D space when a user clicks on individual objects. These menus then allow selections from actions such as renewing or releasing DHCP leases, rebooting individual devices, and changing firewall rules in response to unexpected traffic. The selected actions would be parameterized by information associated with the 3D object's entity.

For example, consider an object with a spin rate indicating what fraction of WAN link capacity was being consumed by the entity's associated device. One of the pull-down menu options might be to instantiate quality of service (QoS) rules that limit the device to one or more preconfigured fractions of the WAN link. End users can focus on the conceptual task rather than the detailed syntax underlying each network or system management action.

Networked Household Devices

The Internet of Things (IoT) envisages a world where everyday devices and systems are augmented with IP-based network interfaces and embedded remote control systems [1]. IoT sees devices such as light switches, washing machines, air conditioning systems, semi-autonomous vacuum cleaners, pacemakers, and in-car navigation and entertainment systems being remotely monitored and/or controlled via their network interfaces.

The "Internet" in IoT may imply connectivity between devices and the public Internet, or it may simply mean a multitude of devices within an organizational domain (e.g., a factory, shopping center, or home) sharing an isolated IP-based communications infrastructure. It is beyond the scope of this article to address the many questions surrounding security and authentication in an IoT world, but Homenet3D offers interesting opportunities therein.

Previously we discussed how Homenet3D supports slave OpenWRT devices feeding their system state updates to a master server, and the master server controlling how each slave's state is mapped into each Homenet3D client's 3D

world. It is easy to generalize the Homenet3D slave to be any network-attached IoT device generating a stream of telemetry data indicating the state of its current environment.

Combine this with custom 3D objects that evoke everyday IoT devices being monitored, flexible placement of all entities with the 3D world, and the potential to trigger scripted actions on slave devices. It is not hard to see Homenet3D being applied to monitoring and controlling a wide range of networked devices around the home from the browser of your smartphone or tablet.

Conclusion

Modern homes see an increasingly complex mix of technologies and devices making up their home network. Usually one IP gateway (but potentially more) sits between the home and broadband service(s) to the outside world. In-home communication involves one or more link layer technologies such as wired Ethernet, 802.11 WiFi or Homeplug (Powerline) networking. Sometimes these technologies are integrated in a single device or spread across multiple devices.

Homenet3D is a project aimed at letting end users observe aspects of their home network infrastructure using increasingly common WebGL-enabled browsers in laptops, tablets, and smartphones. Our goal is to present a useful qualitative view of the home network's current state through the use of suitably designed and animated objects in a virtual 3D environment. To help the broader community explore this in detail we have developed and released an open source BSD-licensed prototype that runs on OpenWRT-based devices. In this article we have discussed our prototype's initial mappings of OpenWRT system state to visually orthogonal attributes of 3D entities viewed by the end user, and Homenet3D's support for customization. Finally, we have outlined ways in which our Homenet3D concept may be relevant to the Internet of Things by providing a user-friendly way to monitor and control a wider range of networked devices.

Acknowledgment

This project has been made possible in part by a gift from the Cisco University Research Program Fund, a corporate advised fund of the Silicon Valley Community Foundation.

References

[1] J. Zheng *et al.*, "The Internet of Things," Guest Editorial, *IEEE Commun. Mag.*, vol. 49, no. 11, Nov. 2011, pp. 30–31.
[2] G. Armitage, "Homenet3D," Sept. 2014; http://caia.swin.edu.au/urp/homenet3d.
[3] T. Chown *et al.*, "IPv6 Home Networking Architecture Principles," IETF Secretariat, Internet-Draft, draft-ietf-homenet-arch-17.txt, July 2014; http://tools.ietf.org/wg/homenet/draft-ietf-homenet-arch/
[4] P. Abel *et al.*, "Automatic Construction of Dynamic 3D Metaphoric Worlds: An Application to Network Management," *Proc. SPIE, Visual Data Exploration and Analysis VII*, vol. 3960, San Jose, CA, Feb 2000, pp. 312–23; http://www.eurecom.fr/publication/276.
[5] D. Chao, "Doom as an Interface for Process Management," *Proc. SIGCHI Conf. Human Factors in Computing Sys.*, ser. CHI '01, Apr. 2001, pp. 152–57; http://doi.acm.org/10.1145/365024.365078.
[6] B. Kot *et al.*, "Information Visualisation Utilising 3D Computer Game Engines Case Study: A Source Code Comprehension Tool," *Proc. 6th ACM SIGCHI New Zealand Chapter's Int'l Conf. Computer-Human Interaction: Making CHI Natural*, ser. CHINZ '05, 2005, pp. 53–60; http://doi.acm.org/10.1145/1073943.1073954
[7] H. Shiravi, A. Shiravi, and A. Ghorbani, "A Survey of Visualization Systems for Network Security," *IEEE Trans. Vis. Comp. Graphics*, vol. 18, no. 8, Aug 2012, pp. 1313–29.
[8] W. Harrop and G. Armitage, "Real-time Collaborative Network Monitoring and Control Using 3D Game Engines for Representation and Interaction," *Proc. 3rd Int'l. Wksp. Visualization for Computer Security*, ser. VizSEC '06, New York, NY, USA: ACM, 2006, pp. 31–40; http://doi.acm.org/10.1145/1179576.1179583.
[9] G. Armitage, "L3DGE — Leveraging 3D Game Engines project," Feb. 2007; http://caia.swin.edu.au/urp/l3dge.
[10] M. Stenberg and S. Barth, "Home Networking Control Protocol," IETF Secretariat, Internet-Draft draft-ietf-homenet-hncp-01.txt, June 2014; http://tools.ietf.org/wg/homenet/draft-ietf-homenet-hncp/.

Biographies

Grenville Armitage (garmitage@swin.edu.au) earned a B.Eng. in electrical engineering (Hons) in 1988 and a Ph.D. in electronic engineering in 1994, both from the University of Melbourne. He is a full professor of telecommunications engineering and founding director of the Center for Advanced Internet Architectures at Swinburne University of Technology. He authored *Quality of Service in IP Networks: Foundations for a Multi-Service Internet* (Macmillan, April 2000) and co-authored *Networking and Online Games — Understanding and Engineering Multi-Player Internet Games* (Wiley, April 2006). He is also a member of ACM and ACM SIGCOMM.

Dominic Allan (domallan8@gmail.com) completed his double degree B.Eng in telecommunications and network engineering and B.Eng in computer science and software engineering in 2014 at Swinburne University of Technology. He worked as a research assistant at Swinburne University's Center for Advanced Internet Architectures in 2013 and 2014 on the Homenet3D project.

Our goal is to present a useful qualitative view of the home network's current state through the use of suitably designed and animated objects in a virtual 3D environment. To help the broader community explore this in detail we have developed and released an open source BSD-licensed prototype that runs on OpenWRT-based devices.

The Greening of Spectrum Sensing: A Minority Game-Based Mechanism Design

Mouna Elmachkour, Essaid Sabir, Abdellatif Kobbane, Jalel Ben-Othman, and Mohammed El koutbi

ABSTRACT

Cognitive radio technology allows the reuse of the underutilized frequency spectrum on an opportunistic and non-interfering basis by means of introducing, besides the legitimate primary users of the spectrum, a new kind of users called cognitive or secondary users. Thus, reliable spectrum sensing is critical to dynamically detect available licensed frequency bands and mitigate the primary signals, but it remains realistically difficult to carry out. In fact, although distributed collaborative sensing has turned out to be fruitful for the cognitive radio environment, its accuracy is often affected by the selfish and autonomous behavior of users. In this article, we model distributed spectrum sensing and channel allocation as a non-cooperative game, and apply the minority game to bring forth and study the cooperative behavior of users. The novelty brought by our study consists of alleviating the number of users contending for primary channels by giving them the opportunity to choose between the two, either sensing the channel or being inactive during the time slot. To address the trade-off faced by the SUs, we evaluate the performance of two secondary systems in a green communications context: energy consumption and transmission delay.

INTRODUCTION

The growth of wireless communication technology (3G, 4G and 5G) and the spread of mobile devices (laptops, PDAs, and smart phones) have widened the gap between spectrum resource and demands. Consumer communication applications such as streaming media, social networks, file downloads, shopping, and Skype require more frequency, latency, and energy efficiency. Furthermore, the radio frequency spectrum as a finite natural resource has become highly mismanaged and strongly underutilized [1].

Cognitive radio (CR) [2, 3] is recognized as the key technology to address the spectrum scarcity and mismanagement problems. In CR networks, secondary users (SUs) are introduced to exploit the licensed frequency bands unused by the primary users (PUs) in time and frequency, as well as in space, under PUs' quality of service (QoS) constraints. Therefore, spectrum sensing is a principal mechanism to detect the available radio frequency bands and decide on accessing the opportunity without interfering with PU signals.

Cognitive devices are expected to instantaneously detect spectral activity of PUs, but in the presence of fading and noise, error-free (perfect) sensing is hard to achieve.

Several techniques and approaches to sensing have been proposed for efficient spectrum sensing, such as distributed and centralized collaborative spectrum sensing [4–6]. With and without central controllers or a dedicated control channel, those approaches suppose a full collaborative exchange of sensing information, taking advantage of diversified received signal strengths. But this scenario is not always realistic considering the selfish and autonomous behavior of users. The spectrum sensing techniques' efficiency is also measured according to their duration and energy consumption [7]. Obviously, cooperation in spectrum sensing and access is claimed to enhance CR network performance, although the sensing information exchange can cause cognitive overload, resulting in deterioration of system performance and intensification of the aggressiveness and non-cooperative behavior rather than cooperation. In this article, we study distributed spectrum sensing and resource allocation in green CR networks from a minority game (MG) theoretic perspective.

The MG [8], as a generalization of El Farol Bar's problem [9], has emerged as a robust tool in modeling cooperation and competition of players given limited resources. In MG, an odd number of players with a set of strategies choose independently between two possible actions: 0 and 1. The players who end up on the minority side win the game and get rewards, while the others lose.

Our main purpose is a sensing decision analysis built around an MG core: each SU has the option to decide whether or not to sense the channel depending on the issue of the implemented game. We compare the impact of SUs'

Mouna Elmachkour, Abdellatif Kobbane, and Mohammed El koutbi are with ENSIAS, Mohammed V University of Rabat.

Essaid Sabir is with ENSEM, Hassan II University of Casablanca.

Jalel Ben-Othman is with the University of Paris.

0163-6804/14/$25.00 © 2014 IEEE

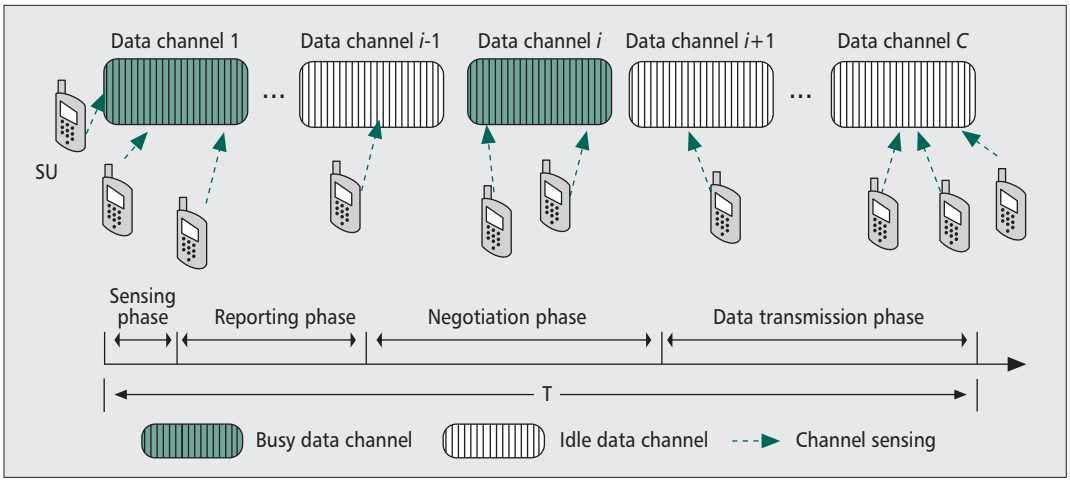


Figure 1. System model. The channel access process consists of four phases. SUs arbitrarily select one data channel to sense at the start of the time slot. According to the channel state, SUs continue through the next phases (reporting, negotiation, and data transmission phase) or leave the process [12].

Consumer communication applications such as streaming media, social networks, file downloads, shopping, and Skype require more frequency, latency, and energy efficiency. Furthermore, the radio frequency spectrum as a finite natural resource has become highly mismanaged and strongly underutilized.

decisions on the average energy consumption and the average transmission delay implementing both original non-cooperative and minority sensing games.

Now, SUs that belong to the minority group have some reward, while the majority group experiences a penalty in the form of regret. When the minority users decide to be inactive for a time slot, they conserve their battery energy.

Indeed, the likelihood of successfully transmitting its data packets refers to the number of active users. When the competition over the spectrum is very high (majority users sense the channel and consequently attempt transmission), the chance of a transmission succeeding becomes low, and thus the risk of spending energy uselessly is high. But when the minority SUs decide to sense the channel, they experience appropriate/comfortable probability of successful transmission. This can be explained by a low number of competitors, and thereafter a low competition over the spectrum. In fact, the non-cooperative MG generates virtual coordination among the SUs without explicit detailed information exchange [10].

MG formulation provides acceptable solutions that might not necessarily be the most optimal. For the non-cooperative approach, SUs that attempt to sense a tagged data channel act selfishly and competitively to get, at the end of the game (the equilibrium), one winner and many losers. However, taking into account the coordination-less feature among the SUs in the MG, it is a tempting approach to solve the coordination issue, especially in highly congested cases. We believe that the MG framework is an attractive and elegant solution to share intelligence among SUs without harming the PUs and without saturating the spectrum with useless signaling messages.

We propose two distributed learning algorithms to converge to both pure Nash equilibrium (NE) and mixed NE for the non-cooperative and MG approaches. Next, we show that the energy consumption in the CR network is minimized under the MG framework.

The balance of this article is as follows. We present the system model and system performance analysis in terms of transmission delay and energy consumption. Next, the non-cooperative game and MG approaches are introduced. We also provide a discussion of the pure NE and fully mixed NE as well. Two distributed learning schemes to display convergence to NE solutions are presented. We provide simulation results illustrating the comparison between the two approaches in terms of delay transmission and energy consumption. Finally, we make concluding remarks.

Spectrum Sensing and Resource Allocation Model

Network Model

We consider a CR network consisting of C data primary channels and one control channel. We assume that channels are slotted in time, and the communication between users is synchronized. M potential SUs are allowed to use the data channels opportunistically without affecting the PUs' communication.

SUs are required to sense the data channels at the start of each time slot. Hence, the data channel state alternates between active state (idle) and inactive state (busy).

We assume that each SU is equipped with at least two transceivers:

- Software-defined radio (SDR) transceiver to sense, receive, and transmit signals.
- A control transceiver to operate over the control channel in obtaining the information of available channels, and negotiating on the corresponding channel.

Similar to [11], we consider that each time slot comprises four phases:

- Sensing
- Reporting
- Negotiation
- Data transmission

In the beginning of a time slot, an SU randomly selects one data channel to sense (Fig. 1). We assume that the C data channels are

If the channel is busy, the station continues checking the channel until it becomes idle. Thus, each node has a transmission probability calculated based on the basic backoff stage and collision probability. When a collision occurs, all nodes back off, and then wait for a random time and retry.

symmetric, that is, channels have the the same QoS and channel occupancy. The sensing phase lasts one mini-slot before the start of the reporting phase on the control channel. The reporting phase is subdivided into C mini-slots, so the reporting for data channel i takes place in mini-slot i. After that, in the negotiation phase, all SUs that sense the ith data channel idle will start negotiation through the carrier sense multiple access with collision avoidance (CSMA/CA) mechanism. At the end of this phase, only one winner starts packet transmission on the channel for the remaining period of the ongoing time slot. In this system model, we prioritize SUs with important data traffic (file transfer, video streaming) to carry on with data transmission for the next time slot if no PU appears. In other words, the data transmission phase begins in the remaining time of the ongoing time slot when channel i is sensed idle, and can continue until the channel becomes busy (i.e., the PU comes back).

Transmission Delay

We deal with transmission delay over y time slots. We assume saturation conditions (i.e., each node always immediately has a packet available for transmission). SUs, once the choice is made to sense the ith channel, use their SDR transceiver to detect the channel state. If a PU signal is detected, SUs will release the process and wait for the next time slot to detect a given ; otherwise, they pass to the reporting phase. During the reporting phase, SUs that choose to sense the ith data channel send beacons on the control channel at the ith mini-slot. Once the number of SUs in contention for channel i, u_i, is known, they pass to the next phase.

The CSMA/CA protocol with request/clear to send (RTS/CTS) mechanism [13] is applied as follows. Each node starts by sensing the channel. If idle, it sends an RTS packet over the channel. When the receiving station detects the RTS, it responds after a short inter frame space (SIFS) with a CTS. If the channel is busy, the station continues checking the channel until it becomes idle. Thus, each node has a transmission probability calculated based on the basic backoff stage and collision probability. When a collision occurs, all nodes back off and then wait for a random time and retry.

The average transmission delay for a given channel is the average useful data transmission duration taking into account the probabilities of collision, successful transmission, negotiation, and also the time duration of the sensing phase, reporting phase, and negotiation phase (considering the negotiation process elucidated above).

Energy Consumption

The proposed resource allocation approach aims to lead SUs to better battery life management through minimal energy consumption. An SU goes through four phases of the process to transmit its packets. Every phase has a minimum required power and time duration that are implied in its energy consumption. We assume that all nodes have the same initial battery capacity. We consider only the power consumed by nodes in sensing, transmitting, receiving, and waiting for the next opportunity.

The total energy consumption for an SU with data packets to transmit on y time slots on data channel i includes the energy consumed in the whole process: power consumed by channel sensing, reporting, negotiation, and data transmission [12].

Channel Sensing and Access Game

We consider a system of M selfish SUs that intend to transmit their data packets over C primary data channels. We assume that at the start of each time slot, each SU randomly selects a single data channel for its data transmission. Let $\Omega_{u_i} = 1, 2, \ldots, u_i$ denote the set of SUs selecting the ith data channel.

As elucidated in the previous sections, the channel access process involves four phases. The channel sensing phase is the first phase; it takes place at the beginning of each time slot and decides on process continuation. We assume that SUs are able to perfectly detect the channel state (idle/busy), although a non-null collision probability with other users (primary or secondary) is inspected in the analysis. If the SU detects the PU signal, and the channel state is busy, the process ends for that user, which will remain inactive until the next slot. If the channel is observed idle, SUs report their sensing information on the control channel by sending a beacon at the mini-slot associated with this data channel. After the reporting phase, the u_i SUs negotiate the data channel for data transmission via CSMA/CA. This phase generates a single winner, while all $u_i - 1$ other contenders leave the process without transmitting their packets. The limited energy has consistently proven to be a serious constraint for SUs through the whole data transmission process; hence, we consider a channel sensing and access game targeted at transmission delay and energy consumption reduction.

Indeed, the channel sensing game involves u_i players who act selfishly through choosing the strategy that maximizes their function payoff. In order to simplify the analysis we consider that all the data channels are sensed by the same number of SUs, that is, $u_i = u_k, \forall i, k \in \{1, \ldots, C\}, i \neq k$.

Each player $j \in \Omega_u$ has two possible actions: to sense the channel or not, that is, $a_j = 1, 0$ to sense and not to sense action, respectively. The payoff function of our game (i.e., the function that represents feedback loss or win of a target player) is related to the two possible actions $a = 1, 0$.

If a target player proceeds to sense the channel, $a = 1$, the payoff will be the difference of the corresponding reward and the spent energy. The reward typically reflects the benefit from the whole process for a player in a given game stage. In our case, the reward $r(a, u^a)$ equals the remaining time to transmit the data packets, $r = T_{Data}$. Obviously, to transmit data packets, a user consumes a part of its energy, which we denote for a given game stage and for a target player as

$E(a, u^a)$, which equals the energy consumed in the channel access process. The payoff function for player j who resolves to start up the process and sense the channel (i.e., $a_j = 1$) can be expressed as p_j. $r(a_j, u^{aj}) - E(a_j, u^{aj})$. p_i is the probability that player j transmits on the tagged channel, and u^{aj} denotes the number of players that select the action $a = a_j$.

If the player chooses not to sense the channel $a = 0$, and thus not lose any energy, nonetheless, the player will lose the chance to transmit its data packets. The payoff function used to evaluate the player j decision $aj = 0$ is represented by a regret function that constantly equals –η, *eta* a positive number.

Now if the player action is $a_j = 1$ and depending on the channel negotiation result, the player could receive a positive payoff f^+ (u^{aj} is the minority group) or a negative payoff f^- (u^{aj} is the majority group), with the assumption that $-\eta > f^-$.

Analysis of the Non-Cooperative Sensing Game

We consider a scenario in which u players are noncooperatively maximizing their utility by updating their actions based on current and common information. For a targeted data channel, at the start of a time slot, u players rationally make their decisions to sense or not to sense the channel at this time slot. Let $\mathbf{a} = \{a_1, a_2, \ldots, a_u\}$ denote a strategy profile for the game. We discuss both the pure strategy and mixed strategy NE for the non-cooperative game $\mathcal{G} = <\Omega_u, \mathbf{a}, \{f_j\}>$. The NE is defined as the set of strategies according to which no player can benefit by unilaterally changing her strategy.

Pure Strategy — A strategy profile $\mathbf{a} = \{a_1, a_2, \ldots, a_{u_i}\}$ is an NE of game $\mathcal{G}$ if for every player j we have

$$f_i(a_j^*, a_{-j}^*) \geq f_i(a_j, a_{-j}^*)$$

where $a_{-j} = a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_{u_i}$. In other words, each player cannot do better by unilaterally changing his strategy.

The game has a unique pure strategy NE.

Mixed Strategy — In the mixed strategy, each player has a probability distribution over the two possible actions, that is, player j can choose to play $a_j = 1$ with probability x_j, and choose $a_j = 0$ with probability $1 - x_j$. We consider a fully mixed strategy in which the probability to select any action is greater than 0. We denote by $\mathbf{x} = (x_1, x_2, \ldots, x_{u_i})$, $0 < x_j < 1 \; \forall_j$, the mixed strategy profile of our game.

A fully mixed strategy NE specifies a fully mixed strategy $x_j^* \in \;]0, 1[$ for each player j (where $j = 1 \ldots u_i$) such that

$$f_j(x_j^*, x_{-j}^*) \geq f_j(x_j, x_{-j}^*)$$

for every fully mixed strategy $x_j \in \;]0, 1[$, where $x_{-i} = x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{u_i}$.

In the equilibrium state, all players choose strategies that maximize their expected payoff.

There is a unique fully mixed NE $\mathbf{x}^*$ that is the solution to $\delta f(\mathbf{x}^*)/\delta \mathbf{x}^* = 0$.

Analysis of the Minority Sensing Game

MG is a non-cooperative game with an odd number of players, where each player has only two alternatives of behavior every turn. The players who end up in the less numerous group (minority group) become the winners.

In the considered resource allocation scheme, after the negotiation phase, only one SU among the u users in contention for one data channel transmits in the data transmission phase. Thus, $u - 1$ SUs, who do not win the contention, lose the energy spent during the previous phases (sensing, reporting, and negotiation). There is a trade-off between sending data packets and conserving battery energy. We consider a traditional MG, that is, the capacity level is $\gamma = (1/2)$ (please see [14] for more details).

At the beginning of a time slot, each player j among u (an odd number) has two strategies: either to sense the channel, 1, or not to sense, 0, that is, $a_j \in \{0, 1\}, j = 1, \ldots, u$.

Therefore, the comfort level of this traditional MG is $(u^1, u^0) = (\Gamma, u - \Gamma)$, with $\Gamma = \lfloor \gamma u \rfloor$.

Let us discus the pure and mixed strategy NE for $\mathcal{G} = <\Omega u, \mathbf{a}, \{f_j\}>$ MG.

Pure Strategy — An NE in pure strategy must satisfy the following two conditions:

$$f(a = 0, \Gamma) \geq f(a = 1, \Gamma + 1)$$

$$f(a = 0, \Gamma - 1) \leq f(a = 1, \Gamma)$$

Hence, no player can do better by unilaterally deviating from the equilibrium.

The NE for pure strategy is when exactly Γ SUs choose to sense the channel, that is, $(u^1, u^0) = (\Gamma, u - \Gamma)$.

Example: NE for a game with three players: (s, n, n), (n, s, n), and (n, n, s) are three possible NE solutions for $u = 3$. We emphasize that there are exactly $\binom{u}{\Gamma}$ asymmetric pure-strategy NEs for the MG.

	s	n
s	(f^-, f^-, f^-)	$(f^-, -\eta, f^-)$
n	$(-\eta, f^-, f^-)$	$(-\eta, -\eta, f^+)^*$

	s	n
s	$(f^-, f^-, -\alpha)$	$(f^+, -\eta, -\eta)^*$
n	$(-\eta, f^+, -\eta)^*$	$(-\eta, -\eta, -\eta)$

Mixed Strategy — In the mixed strategy, each player has a probability distribution over the two possible actions; that is, player j can choose to play $a = 1$ with probability x_j, and choose $a = 0$ with probability $1 - x_j$. We consider a fully mixed strategy in which the probability of selecting any action is greater than 0. We denote by $\mathbf{x} = (x_1, x_2, \ldots, x_u)$, $0 < x_j < 1 \; \forall j$, the mixed strategy profile of our game.

Players modify their previous strategies with some probability if they learn that they would get a higher payoff if playing differently. This decision is performed through a learning rule, which leads to the algorithm converging (or not) toward the equilibrium solution.

A fully mixed strategy Nash equilibrium for each player j is such that

$$f_j(x_j^*, x_{-j}^*) \geq f_j(x_j, x_{-j}^*)$$

for every fully mixed strategy $x_j \in]0, 1[$, where $x_{-j} = x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{u_i}$. In the equilibrium state, all players choose strategies that maximize their expected payoff. There exists a unique fully mixed Nash equilibrium $\mathbf{x}^*$ that is the solution to

$$\frac{\delta f(\mathbf{x}^*)}{\delta \mathbf{x}^*} = 0.$$

Distributed Learning NE solutions

Distributed learning algorithms entail a sequence of plays of a static "one-shot" players. Each player would adjust its strategy based on its local observations to maximize its reward. Indeed, players modify their previous strategies with some probability if they learn that they would get a higher payoff if playing differently. This decision is performed through a learning rule, which induces the algorithm converging (or not converging) toward the equilibrium solution.

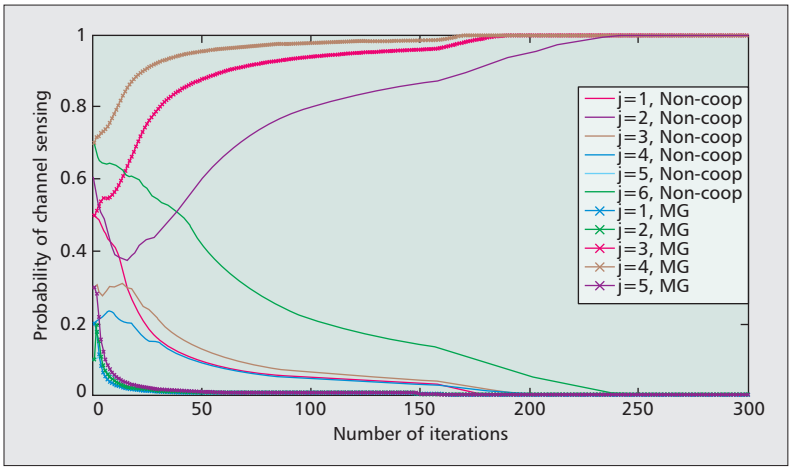


Figure 2. Pure strategy NE convergence: non-cooperative game and MG.

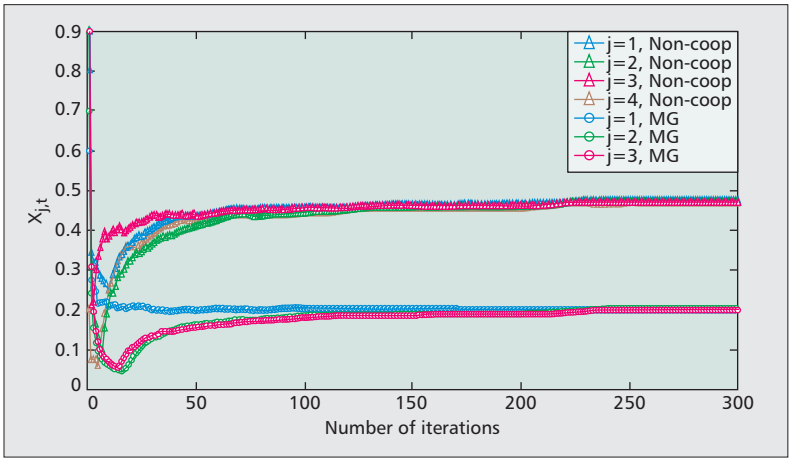


Figure 3. Mixed strategy NE convergence: non-cooperative game and MG.

Pure Strategy NE Convergence

In our case, we use the Linear Reward Inaction algorithm (LRI) [15] to reach the NEs characterized in the section above for the non-cooperative game and the MG. This algorithm converges to one of the pure strategy coordination equilibria while the initial population distributions are asymmetric.

Initially, each SU i picks an action (to sense or not) based on the initial distribution $x(0)$. At each round, the players decide on the action that maximizes their payoff through the LRI updating rule.

For the non-cooperative sensing game, at each game step, u SUs select a tagged data channel for their data transmission. To sense the data channel is the first step before reporting, negotiating, and finally transmitting the data packets. In the non-cooperative sensing game considered, the u players compete with each other selfishly to sense the channel. Based on its action, the actions of other players, and the channel state (idle/busy, channel utilization rate, collision probability), each player attempts to maximize its payoff and be the only transmitter on the data channel. Figure 2 shows that the pure NE learning iterations lead to a unique game winner, player 2 for this round, of six players. Another execution of the learning algorithm might give another winner, but always a unique winner that can sense the channel. This means that only one SU starts the process of transmission. Other players remain inactive for the ongoing time slot and will repeat the sensing game for other data channels in upcoming time slots.

In the MG, in each game step, the u SUs in competition for channel sensing attempt to belong to the minority with the aim of a better chance to transmit their data and conserve their energy. Figure 2 illustrates the pure NE learning of MG between five players. As elucidated above, $\lfloor u/2 \rfloor$ players constitute the minority, the players 3 and 4 in this example. The three other players will repeat the game of sensing for another data channel in the next time slot.

Mixed Strategy NE Convergence

For the mixed strategy NE, we have opted for the Gradient Descent algorithm. This algorithm involves that the players update their beliefs at each step, choosing the best response to the new action profile by computing a gradient response. Figure 3 shows the mixed strategy NE convergence for the non-cooperative game and MG. For a set of four players, the probability of opting for channel sensing considering a non-cooperative sensing approach converges to symmetric NEs, ($x_{j,t} = 0.45, j = 1 \ldots 4$). For the MG sensing approach, we evaluate the NE convergence for three players. The Gradient Descent learning algorithm for our game converges to a symmetric NE, ($x_{j,t} = 0.22, j = 1 \ldots 3$).

Numerical Results

In this section, we evaluate the energy consumption and transmission delay performance of a cognitive network comparing the two approaches: non-cooperative and MG sensing. SUs are allowed

to decide at the beginning of the time slot whether to sense the channel or to be inactive. Thus, an SU focuses on the trade-off between energy consumption and data transmission.

We consider $u = 7, 13$, and compare the system performance with the two approaches (the restriction of an odd number of players for the MG). Figure 4 shows the average energy consumption as a function of the number of time slots required to achieve data transmission y, and the number of SUs in competition for channel u. The average energy consumption decreases with the increase of y and u for the two approaches. The energy consumption for the non-cooperative approach is considerably more important than for the MG approach. The average transmission delay is also important for the non-cooperative game, as depicted in Fig. 5. This shows that SUs can significantly preserve their battery energy and optimize the data transmission delay through the MG approach, due to the fact that players in MG cooperate with each other with no intention of cooperating.

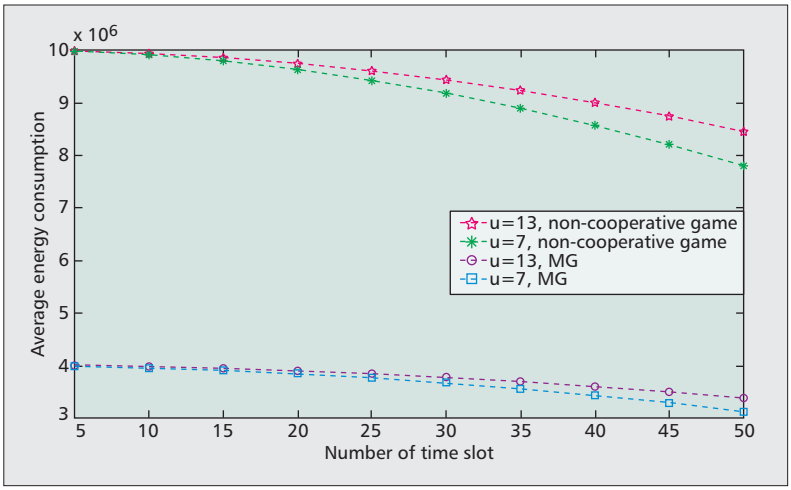


Figure 4. Average energy consumption: non-cooperative sensing/MG sensing.

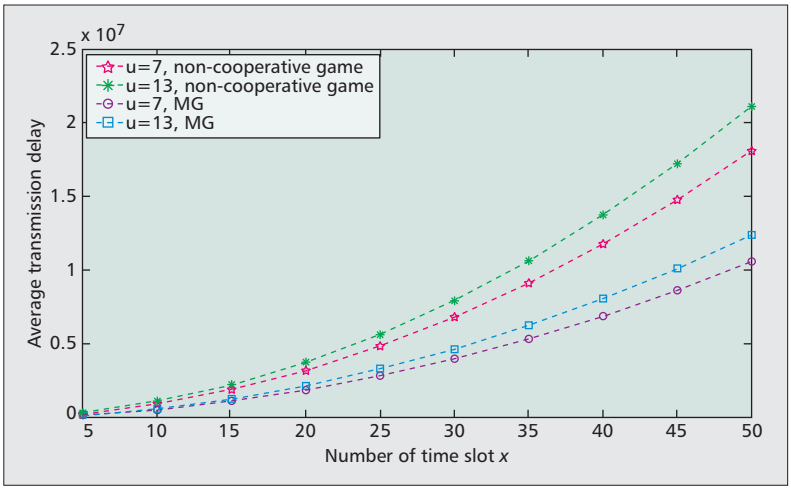


Figure 5. Average transmission delay: non-cooperative sensing/ MG sensing.

Conclusion

In this article, we propose a distributed spectrum sensing and channel access scheme for a cognitive network based on an original non-cooperative game and on a minority game modeling the interaction among autonomous and selfish SUs. We point out the impact of the potential number of SUs in competition for a targeted channel on average energy consumption and average transmission delay. We formulate a payoff function comprising the transmission delay and energy consumption related to user strategies. Then we analyze strategic behavior of SUs in primary channel contention and study the NE solution of the sensing game for both pure and fully mixed strategies. Moreover, we use distributed learning algorithms that leads SUs to attain their Nash optimal strategies. Finally, we show through simulations that cognitive users' battery life and transmission delay improve through MG-based distributed spectrum sensing and channel access scheme. Our ongoing work consists of improvement of the spectrum sensing mechanism. We aim to use the concept of a coalitioanal game to organize the secondary users into channel sensing coalition structures considering their battery levels and data traffic distinction.

References

[1] FCC, "Facilitating Opportunities for Flexible, Efficient, and Reliable Spectrum use Employing Cognitive Radio Technologies," tech. rep. ET Docket no. 03-108, 2003.

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

[3] J. Mitola, *Cognitive Radio Architecture: The Engineering Foundations of Radio XML*, Wiley, 2006.

[4] S. Mishra, A. Sahai, R. Brodersen, "Cooperative Sensing among Cognitive Radios," *Proc. IEEE ICC*, Turkey, vol. 4, 2006, pp. 1658–63.

[5] F. Chen and R. Qiu, "Centralized and Distributed Spectrum Sensing System Models Performance Analysis Based on Three Users," *Proc. Int'l Conf. Wireless Commun. Networking and Mobile Computing*, Chengdu, China, 2010, pp. 1–4.

[6] S. Maharjan *et al.*, " Distributed Spectrum Sensing in Cognitive Radio Networks with Fairness Consideration: Efficiency of Correlated Equilibrium," *Proc. IEEE Int'l. Conf. Mobile Adh Hoc and Sensor Systems (MASS*, Valencia, Spain, 2011, pp. 540–49.

[7] S. Choi S and K. G. Shin, "Secure Cooperative Spectrum Sensing in Cognitive Radio Networks Using Interference Signatures," *Proc. IEEE Conf. Commun. and Network Security*, Washington, DC , 2013, pp. 19–27.

[8] D. Challet and Y. C. Zhang, "Emergence of Cooperation and Organization in an Evolutionary Game," *Physica A: Statistical Mechanics and Its Applications*, 1997, vol. 246, no. 3, pp. 407–18.

[9] W. B. Arthur, "Inductive Reasoning and Bounded Rationality, (The El Farol Problem)" *American Economic Review*, 1994, vol. 84, no. 2, pp. 406–11.

[10] P. Mahonen and M. Petrova, "Minority Game for Cognitive Radios: Cooperating Without Cooperation," *PHYSCOM* (Elsevier), 2008, vol. 1, no. 2, pp. 94–102.

[11] M. Elmachkour *et al.*, " New Insights from a Delay Analysis for Cognitive Radio Networks with and without Reservation," *Proc. IWCMC'12*, 2012, Limassol, Cyprus, pp. 65–70.

[12] K. Xu, M. Gerla, and S. Bae, "How Effective is the IEEE 802.11 RTS/CTS Handshake in Ad Hoc Networks?," *Proc. IEEE GLOBECOM*, Taipei, Taiwan, 2002, pp. 72–76.

[13] M. Elmachkour et al., "Data Traffic-Based Analysis of Delay and Energy Consumption in Cognitive Radio Networks with and without Resource Reservation," *Int'l. J. Commun. Systems*, 2014.

[14] M. Elmachkour *et al.*, "Green Opportunistic Access for Cognitive Radio Networks: A Minority Game Approach," *Proc. IEEE ICC*, Sydney, Australia, 2014, pp. 5372–77.

[15] M. A. L. Thathachar , P. S. Sastry, and V. V. Phansalkar, "Decentralized Learning of Nash Equilibria in Multiperson Stochastic Games with Incomplete Information," *IEEE Trans. Sys. Man Cyber.*, 1994, vol. 24, no. 5, pp. 769–77.

BIOGRAPHIES

MOUNA ELMACHKOUR (mouna.elmachkour@gmail.com) received her B.Sc. degree in mathematics and computer sciences (2008) from Mohammed V University, Rabat, Morocco, and her M.S in management of information systems and computer communications (networks) in 2010, from University Mohammed Benabdellah, Faculty of Sciences, Fez, Morocco. She is currently a Ph.D student in ENSIAS, Rabat, Morocco. Her research interests include ad hoc networks, cognitive radio networks, DTNs, system performance evaluation, and game theory.

ESSAID SABIR [SM] received his B.Sc. degree in electrical engineering electronics and automation (2004) from Mohammed V University and his M.Sc in Telecommunications and Wireless Engineering (2007) from th National Institute of Post and Telecommunications, Rabat, Morocco. In 2010, he received his Ph.D degree in networking and computer sciences jointly from the University of Avignon, France, and Mohammed V University. He served as a contractual associate professor at the University of Avignon from 2009 to 2012. He serves as a reviewer for prestigious international journals (Springer-WINET, Elsevier-COMNET/COMCOM/IJEC, Wiley-JWCMC, JCDS,etc.) and international conferences (IEEE GLOBECOM, ICC, WCNC, ICT, IWCMC, WIOPT, etc.). He is or has been involved in several national and international/European projects. Currently, he is a full-time assistant professor at the National Higher School of Electricity and Mechanics (ENSEM). His current research interests include protocols design, ad hoc networking, cognitive radio, stochastic learning, networking games, pricing and network neutrality. He has co-authored over 15 journal articles, one book, two book chapters, and over 40 conference publications. He was the recipient of the best paper award at the IEEE International Conference on Next Generation Networks and Services (2014), and has been nominated at many other events. He received the Exchange Grant of the Center of Excellence in ICT, funded by INRIA-France (2007–2010), offered every year to the top three Moroccan Ph.D. candidates. He also was a recipient of the graduate scholarship (2007–2010) from the National Centre for Scientific and Technical Research, Morocco. He is a founder and Vice-Secretary General of the Moroccan Mobile Computing and Intelligent Embedded-Systems Society (Mobitic). As an attempt to bridge the gap between academia and industry, he has founded the International Workshop on Ubiquitous Networking (UNet) and co-founded the International Workshop on Wireless Networks and Mobile Communications (WINCOM), a successful event recently converted to an international conference.

ABDELLATIF KOBBANE [M] has been an associate professor at the Ecole Nationale Suprieure d'Informatique et d'Analyse des Systemes (http://www.ensias.ma/, ENSIAS), Mohammed V University (http://www.um5s.ac.ma/) Mobile Intelligent System (MIS) Team since 2009. He received his Ph.D. degree in computer science from Mohammed V-Agdal University, Morocco, and the University of Avignon in September 2008. He received his research M.S. degree in computer science, telecommunication, and multimedia from HYPER-Mohammed V-Agdal University in 2003. His research interests lie with the field of wireless networking, performance evaluation in wireless networks and NGN, ad hoc networks, DTNs, mesh networks, cognitive radio, mobile computing, mobile social networks, MTC and M2M systems, heterogeneous networks, and future networks. His work appears in highly respected international journals and conferences, including IEEE ICC, GLOBECOM, IWCMC, WCNC, Networking, and PIMRC. He has supervised and co-supervised several graduate students in these areas. He is widely known for his work on wireless networks, in particular cognitive radio and MTC networks. He has more than 10 years of computer sciences and telecom experience, in Europe (France) and Morocco, in the areas of performance evaluation in wireless mobile networks, mobile cloud networking, cognitive radio, ad hoc networks, and future networks. He has served as a reviewer for many international journals and conferences such as COMCOM, ICC, GLOBECOM, IWCMC, ICNC, and WCNC. He is founder and President of the Association of Research in Mobile Wireless Networks and Embedded Systems (MobiTic) in Morocco, and Co-Chair of the Wireless Networking Symposium of the 10th and 11th International Wireless Communications & Mobile Computing Conferences (IWCMC '14–15). He was TCP Chair of the 5th and 4th Workshop on Codes, Cryptography and Communication Systems (WCCCS). He served as Local Chair for the 5th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WIMOB 2009). He has also served as an Invited Speaker at ENSI (Tunis), with his speech, "Games Theory Applied to Wireless Networks," November 13, 2013, Hammamat, Tunis, and INRIA, France, with a speech on "Performance Evaluations in Wireless Communications Networks," September 19, 2013.

JALEL BEN-OTHMAN [SM] received his B.Sc. and M.Sc. degrees, both in computer science, from the University of Pierre et Marie Curie, (Paris 6), France, in 1992 and 1994, respectively. He received his Ph.D. degree from the University of Versailles, France, in 1998. He was an assistant professor at the University of Orsay (Paris 11) and Paris 6 in 1998 and 1999, respectively. He was an associate professor at the University of Versailles from 2000 to 2011. He has been a full professor at the University of Paris 13 since 2011. His research interests are in the areas of wireless ad hoc and sensor networks, broadband wireless networks, multi-services bandwidth management in WLAN (IEEE 802.11), WMAN (IEEE 802.16), WWAN (LTE), VANETS, sensor and ad hoc networks, vehicular networks, Internet of Things, and security in wireless networks in general and wireless sensor and ad hoc networks in particular. His work appears in highly respected international journals and conferences, including IEEE ICC, GLOBECOM, LCN, MSWIM, VTC, and PIMRC. He has supervised and co-supervised several graduate students in these areas. He is an editorial board member of Elsevier *Computer Networks* (COMNET), Wiley *Wireless Communications and Mobile Computing* (WCMC), Wiley *Security and Communication Networks* (SCN), Interscience *International Journal of Satellite Communications Policy and Management*, *IEEE Journal of Communications and Networks*, *International Journal of Information and Communication Technology*, and *International Journal on Advances in Networks and Services*. He is also an Associate Editor of Wiley *International Journal of Communication Systems*. He has served as a member of Technical Committees of more than 100 international IEEE/ACM conferences and workshops including ICC, GLOBECOM, MSWIM, LCN, and MASCOTS. He is a member of ACM. He served as Local Arrangement Chair for the 13th IEEE International Symposium on Computer Communication, 2009. He served as a TPC Co-Chair of IEEE GLOBECOM Wireless Communications Symposium (GLOBECOM 2010) and 9th and 10th International Workshops on Wireless Local Networks (2009 and 2010). He served as a publicity chair of several conferences such as the 12th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks '10, and the 25th Biennial Symposium on Communications. He has served as TPC Co-Chair for IEEE GLOBECOM Ad Hoc and Sensor and Mesh Networking (GLOBECOM 2011, 2014), 6th ACM International Symposium on QoS and Security for Wireless and Mobile Networks (2010, 2011, 2012), Wireless Networking Symposium of the International Wireless Communications and Mobile Computing Conference (2011, 2012, 2013, 2014, 2015), IEEE International Conference on Communications Ad Hoc and Sensor and and Mesh Networking (ICC 2012, ICC 2014). He has served for other conferences such as ICNC, WSCP, CNIT. He has also served as Tutorial Chair for Modeling, Analysis and Simulation of Computer Telecommunication Systems (MASCOTS 2014). He was the Secretary and currently Vice Chair of the IEEE Ad Hoc and Sensor Networks Technical Committee since January 2012. He is an active member of IEEE CIS-TC, Communication Software, and WTC.

MOHAMMED EL KOUTBI obtained his engineering degree in computer science from Ecole Mohammadia d'Ingénieurs (higher engineer school at Rabat) in 1988. He obtained his Ph.D. in computer science from Montreal University, Canada, in July 2000. He works now as a full professor at ENSIAS, Mohammed V University. He is very active in the domain of ad hoc routing protocols. He also does some research in the field of model transformations based on UML. He is the head of the Mobile Intelligent Systems research team at ENSIAS.

Video Adaptation for Consumer Devices: Opportunities and Challenges Offered by New Standards

James Nightingale, Qi Wang, Christos Grecos, and Sergio Goma

ABSTRACT

Video and multimedia streaming services continue to grow in popularity and are rapidly becoming the largest consumers of network capacity in both fixed and mobile networks. In this article we discuss the latest advances in video compression technology and demonstrate their potential to improve service quality for consumers while reducing bandwidth consumption. Our study focuses on the adaptation of scalable, highly compressed video streams to meet the resource constraints of a wide range of portable consumer devices in mobile environments. Exploring SHVC, the scalable extension to the recently standardized High Efficiency Video Coding scheme, we show the bandwidth savings that can be achieved over current encoding schemes and highlight the challenges that lie ahead in realizing a deployable and user-centric system.

INTRODUCTION

Advances in portable consumer devices such as smartphones have boosted global popularity of high-quality multimedia-rich applications over WiFi, 3G, and 4G wireless networks. Cisco [1] has predicted that over half of all Internet traffic will be generated by mobile devices by 2017, with a remarkable 13-fold increase from 2012, and the vast majority of the traffic is expected to be video related. Though substantial volumes of mobile traffic are currently offloaded to wired networks via WiFi connections, the trend is toward ubiquitous content consumption over all available wireless mediums. This rapid growth, while offering myriad opportunities for content providers and advertisers, will place huge burdens on the network infrastructure and pose significant challenges for Internet Service Providers (ISPs).

The main implication associated with this unprecedented traffic growth is the demand for more advanced management of network capacity and application bandwidth requirements, while meeting consumers' expectations. Users expect to receive high-quality services by fully exploiting the capabilities of their devices. Given the massive improvements in hardware over the last few years, with High-Definition (HD) or better displays and quad-core processors becoming the norm for even modestly priced handsets, this is a daunting task. Service providers will need to deliver enormous volumes of delay-sensitive traffic, possibly over imperfect wireless channels, to a plethora of portable devices with a wide variety of processing and playback capabilities, at the same time, ensuring that each customer's experience is maximized and their loyalty retained.

In this article we highlight three separate but intricately woven challenges. First, by improving video compression techniques, bandwidth requirements can be reduced without compromising on video quality. Second, employing scalable and adaptive video streaming technologies, network resources can be conserved by matching video stream characteristics to user device capabilities. Third, video streams can be dynamically and intelligently adapted to the varying status of the transmission channels.

With the standardization of High Efficiency Video Coding (HEVC) [2] as H.265 by the ITU-T and as 23008-2 MPEG-H Part 2 by ISO/IEC in 2013, a major step forward has been taken in delivering high-quality video streams at greatly reduced bandwidths. Bandwidth reductions of up to 50 percent can be achieved using HEVC, when compared with current video compression standards, without loss of picture quality. HEVC is an effective means to address bandwidth consumption, the first of the three highlighted challenges.

SHVC [3], a more recent and as yet incomplete, scalable extension to HEVC, will offer the facility to adapt an HEVC-encoded video stream to meet end user device capabilities and network resource constraints. By offering a single encoded bitstream containing multiple representations of the same video at different spatial, temporal, or quality resolutions, SHVC can be adapted, statically or dynamically, to match a device constraint such as the display resolution. Using SHVC will address the second of the three highlighted challenges by conserving bandwidth, as only those sub-streams that the client device is capable of playing will be sent over the network.

James Nightingale, Qi Wang, and Christos Grecos are with the University of the West of Scotland.

Sergio Goma is with Qualcomm Inc.

0163-6804/14/$25.00

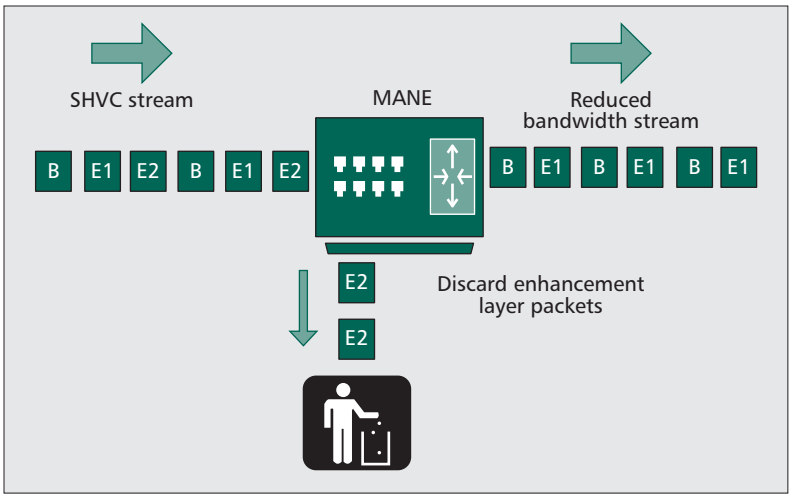


Figure 1. Adapting an SHVC scalable video stream at a MANE to reduce the bandwidth requirement.

In the third and most demanding of the challenges, SHVC adaptation solves the problem of meeting prevailing network constraints by only transmitting the number of sub-streams or layers that can be successfully delivered under the current network conditions. Adaptation of scalable video streams takes place at a Media Aware Network Entity (MANE) located within the network. The principle of scalable video stream adaptation to meet a bandwidth restriction is shown in Fig. 1. In the illustration a video stream consists of three layers (Base, Enhancement 1, and Enhancement 2). A MANE reduces the bandwidth of the stream by discarding packets from the highest enhancement layer (E2). If further reduction is required, the lower enhancement layer (E1) would also be discarded. Transmitting only the base layer will deliver a lower visual quality.

Matching video streams to network capabilities helps to prevent issues such as buffering waits or heavily distorted pictures, which would affect a user's perception of quality.

This article provides both valuable insights into video stream adaptation using the new compression technologies of HEVC and SHVC, and the results of a comparative study (between current and evolving scalable video encoding schemes) conducted on a wireless networking platform. The relative merits of competing adaptive streaming technologies are discussed primarily within the context of mobile/portable devices, and an outline model of an SHVC-based streaming system is presented. Results of an experimental implementation of this model show that bandwidth savings of between 42 percent and 58 percent can be achieved using SHVC without any significant quality penalty. The limitations of the current SHVC design are also explored by evaluating its performance under adverse network conditions.

High Efficiency Video Coding

In 2010 the Joint Collaborative Team on Video Coding (JCT-VC) was formed with a goal of designing a new generation of video encoding standards that would offer a 50 percent bandwidth saving over the H.264 Advanced Video Coding (H.264/AVC) standard. In addition to improving coding efficiency, the JCT-VC also sought to deliver a standard that would improve integration with transport system protocols, include resilience to data loss, and be suitable for implementation on parallel processing architectures. Another important consideration in the codec design was the need to address the encoding and transmission requirements of emerging Ultra-High Definition Television (UHDTV) formats such as 4K (2160p) and 8K (4320p) with spatial resolutions of up to 16 times that of current HD (1080p) displays.

Target applications for the new standard included broadcast television at HD and higher resolutions over cable, satellite, and terrestrial transmission channels, video delivery over Internet and mobile networks, video-conferencing, and storage applications such as blue-ray discs. Over a period of almost three years many encoding tools, proposed mainly by industry, were evaluated, culminating in the adoption of HEVC as a new standard for video coding. HEVC retains the same two-layer hybrid coding structure as its predecessor (H.264/AVC), in which a Video Coding Layer (VCL) handles encoding and a Network Abstraction Layer (NAL) organises the HEVC-encoded bitstream as a series of logical data units (NAL units) for encapsulation and transmission.

Adaptive Video Streaming

Adapting video streams to meet device and network constraints has been the subject of extensive research in recent years. Two distinct approaches have evolved to deliver video streams to a diverse range of devices over fluctuating network channels.

DASH

Dynamic Adaptive Streaming over HTTP (DASH) [4] has recently become popular with content providers. In DASH, several representations (typically three) of the same video, each with a different bitrate, are either separately encoded or derived by transcoding, up-sampling, or down-sampling from a single representation. Each representation, which is contained in a separate file, is divided into segments marked in relation to their temporal position within the video sequence. The client device monitors the arrival of segments from the server over short time intervals. Using its knowledge of the current delivery rate and the bandwidth requirements for each representation, it selects which representation to request segments from during the next time interval. Switching between representations can take place frequently during streaming sessions where bandwidths fluctuate or losses and delays in the network lead to the need for retransmissions.

H.264/SVC

The alternative approach is to employ a single video file containing a number of layers that can be extracted, either independently or in combination, to match a given constraint (such as

bitrate, spatial resolution, etc.). This approach has previously been adopted as the scalable extension (H.264/SVC) [5] to the H.264/AVC standard. A similar extension (SHVC) is currently being developed by the JCT-VC for HEVC. H.264/SVC permits encoding of a video stream consisting of an H.264/AVC-compliant base layer, ensuring compatibility with legacy decoders, and a number of enhancement layers.

Three possible scalable dimensions (spatial, temporal, and quality) are available in H.264/SVC. Spatial scalability offers a course-grained mechanism whereby the H.264/AVC-compliant base layer is encoded at a lower spatial resolution with a spatial enhancement layer carrying the additional encoded picture data required to increase the number of pixels by a factor of four. For example, base layer = Height (H) × Width (W), and base layer + enhancement layer = 2H × 2W.

In temporal scalability, the base layer is encoded at a low frame rate with each enhancement layer increasing the frame rate by a factor of two. For instance, base layer = 7.5 frames per second (f/s), enhancement layer 1 = 15 f/s, and enhancement layer 2 = 30 f/s. Temporal scalability can be combined with spatial scalability with each spatial layer comprising a number of temporal layers.

Quality scalability alters the Quantization Parameter (QP) used by the encoder for each quality layer in a bitstream. A high QP will produce a more condensed bitstream, encoded using fewer bits but with a higher level of error (noise) in the resultant picture. In quality scalability the highest QP is chosen when encoding the base layer. Enhancement layers increase quality by reducing the QP, resulting in higher bitrates and reduced noise in the picture.

All three scalable dimensions can be combined to produce very granular scalable bitstreams containing many layers. Examples of each type of scalability are shown in Fig. 2. H.264/SVC scalable video coding lowers coding efficiency by up to 20 percent when compared with H.264/AVC [5].

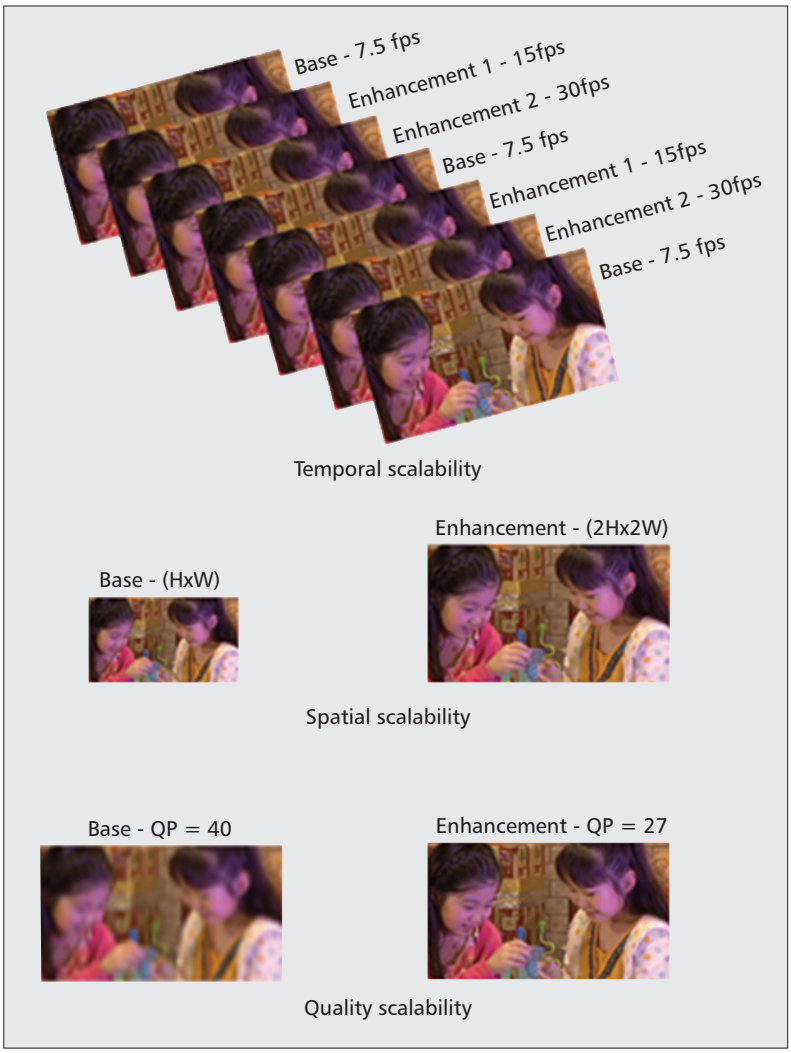


Figure 2. The three scalable modes. Temporal (frame rate) scalability is achieved by removing frames belonging to enhancement layers. In spatial (picture resolution) scalability the base layer has a lower spatial resolution and in quality scalability the base layer has a 'noisier' picture than an enhancement layer.

SHVC

SHVC has not yet reached the same level of maturity as H.264/SVC. The current working model only supports a two-layer scalable bitstream (base layer and a single enhancement layer). In addition to the three scalability dimensions employed in H.264/SVC, hybrid scalability, Region Of Interest (ROI) scalability, and color gamut scalability are all actively being considered for inclusion in the SHVC specification.

Temporal scalability has been incorporated into the HEVC design, allowing a specified number of temporal layers to be extracted as a standard-compliant bitstream. Similarly to H.264/SVC, SHVC quality scalability is achieved by varying the QP value of each layer. Spatial scalability in SHVC has two modes. The resolution of an SHVC spatial enhancement layer can be either twice the height and width of the base layer (2× mode) or one and a half times the height and width of the base layer (1.5× mode).

The first new scalability mode being considered for SHVC is hybrid scalability where the base layer complies with the H.264/AVC standard and the single enhancement layer is encoded by employing HEVC. In this hybrid mode, legacy H.264/AVC decoders will still be able to decode the base layer. Example applications for this mode may include broadcast transmission of television programs at greatly increased spatial resolutions such as UHDTV. Legacy TVs and set top boxes would be able to decode the base layer at the current HD quality, while new devices would decode the entire UHDTV stream. Some of the current proposals [6] being considered for the hybrid mode include having an HEVC-encoded temporal layer of the H.264/AVC base layer and an enhancement layer that offers combined scalability (temporal and spatial). The latter proposal would allow the enhancement layer to transmit streams with both a higher spatial resolution and an increased frame rate.

ROI scalability is another interesting proposal for SHVC, where the enhancement layer carries a higher quality representation of a selected

Although SHVC is still under development and no SHVC-specific MANE implementations have been reported, models previously proposed for H.264/SVC can be readily updated to utilize the new scalable encoding format.

region within the viewable area. There are many potential applications for ROI scalability in, for example, security surveillance and sports broadcasting. For security applications, an observer may select an area (or object) to enhance. This area would then have additional picture information transmitted in the enhancement layer to perhaps reduce the noise (quality scalability) in the selected region. This method offers the advantage of only carrying an additional bitrate burden for the chosen region of the picture in the enhancement layer, thereby saving bandwidth and reducing congestion in transmission channels.

Color gamut scalability recognizes that current HD standards (ITU-R Rec. BT 709) employ an 8-bit color encoding depth, whereas UHDTV applications will utilize a 10-bit or 12-bit color depth (ITU-R Rec. BT 2020). The aim of this scalability vector mode in HEVC is to permit a base layer that is encoded with an 8-bit color depth and an enhancement layer using 10-bit or 12-bit color depth. Legacy decoders would be able to decode the base layer only. This scalability mode is seen as a priority for the implementation of HEVC in broadcast transmission systems.

SHVC vs. DASH

Storage Requirements

Although both DASH and scalable video encoders such as H.264/SVC and SHVC address the same issues of stream adaptation to meet a set of resource constraints, they operate in significantly different ways. DASH requires that multiple representations of the same content are individually stored on the server, while scalable encoding requires only a single file. It is noted that although employing SHVC encoding reduces compression efficiency compared with HEVC in the range from 8 percent to 22 percent, dependent on encoding mode [7], SHVC layered bitstream files are actually significantly smaller than the sum of the multiple representations required by DASH. Disk space savings of 35 percent to 40 percent were observed in SHVC when compared with DASH (based on HEVC encoding) in our empirical studies. Hardware costs of content delivery networks, which typically have several copies of popular content distributed across the network, could be reduced by utilizing scalable encoding methods.

Flexibility

A second consideration when comparing DASH with scalable video encoding is the location within an end-to-end network path where video stream adaptation occurs. DASH-based streaming is controlled by the client device responding to changes in bandwidth and segment delivery. The choice of which representation (quality level) to request from the server during the next time interval is derived from delivery reports for the current time interval, whereas in SHVC and other scalable video streams, adaptation occurs at a MANE, which can conceptually be located at either the end point(s) of a transmission path or intermediate nodes such as media gateways or proxies. Network packets carrying SHVC bitstreams can be easily parsed with low complexity at intermediate nodes to determine the scalable layer identity of the packet contents. This approach is more flexible than that of DASH as it allows network operators to apply global traffic management policies to video streams traversing the network. DASH does, however, offer advantages over SHVC and H.264/SVC in that HTTP traffic can readily traverse firewalls and NAT access controls, and adaptation is performed at the client without the need for a MANE at the network side. The third, and less well researched, consideration when comparing DASH and scalable video streaming is that of a user's perceptual Quality of Experience (QoE) when delivery takes place over an imperfect transmission channel. Again SHVC and other scalable encoding schemes offer more flexibility in how they can be implemented than DASH does. SHVC-based streaming can be implemented over either reliable (e.g. TCP) or unreliable (e.g. UDP) transmission protocols to accommodate different use case scenarios. For example, it may be more appropriate to have an unreliable transport protocol with channel coding (e.g. Forward Error Correction (FEC)) for packet loss mitigation and SHVC-based adaptation in some real-time streaming scenarios where packet loss or interference in wireless media is an issue. In such a scenario reliable delivery mechanisms, including DASH, may experience significant buffering delays [8] due to large numbers of retransmissions. In contrast, SHVC-based schemes with error and loss compensation will deliver video content, although at a lower perceptual quality, while minimizing buffering waits. Objective comparisons between sent and received files using, for example, the Peak Signal to Noise Ratio (PSNR) metric may well score the reliable transport stream delivery higher than the unreliably delivered stream. However, in a subjective evaluation, viewers may find the buffering wait intolerable. Current standard methods of video quality assessment do not accurately reflect the detrimental effect of buffering on a user's QoE. The Video Quality Experts Group (VQEG) is currently considering new methods of subjective assessment for adaptive streaming mechanisms such as DASH, in which additional factors related to a user's perception of quality will be measured [9].

SHVC Pilot Implementation

Although SHVC is still under development and no SHVC-specific MANE implementations have been reported, models previously proposed for H.264/SVC can be readily updated to utilize the new scalable encoding format. The SHVC evaluation described here is implemented by building upon the H.264/SVC streaming framework proposed by Nightingale, Wang, and Grecos in [10].

The SHVC MANE is implemented as a software agent co-located and fully integrated with the streaming server. The combined streaming server and MANE consist of three principal components: network resource matching, scheduling, and dependency checking. Network resource matching ensures that only those layers that can be successfully delivered by the currently available bandwidth are transmitted. On a

Group Of Pictures (GOP) basis, the available bandwidth during the transmission period of the GOP is compared with the cumulative encoded bandwidth of all the layers in the scalable SHVC bitstream. Where the encoded bitrate exceeds the available bandwidth the enhancement layers are dropped recursively, beginning with the highest enhancement layer, until the encoded bitrate of the remaining layers is less than or equal to the available bandwidth. Our experimental evaluation employs a software agent located within the Wide Area Network (WAN) emulation nodes (Fig. 3) to inform the MANE of currently available bandwidth and end-to-end delay.

Conceptually, this monitoring and feedback mechanism could be placed at other point(s) in the network, including at the client device. Second, delays in the transmission network are accommodated by ensuring that only those packets anticipated to arrive at the client in time to be decoded and displayed without causing buffering waits are transmitted. This is achieved by permitting an initial start-up delay (buffering window) of 150ms at the client. The arrival time of each packet is estimated relative to the first packet in the stream. Only those packets estimated to arrive within the buffering window are transmitted. Finally, as with all recent video encoding standards, some frames are predicted from others and cannot be successfully decoded unless the frame(s) upon which they are dependent are available at the decoder. To conserve bandwidth, a dependency checking mechanism is employed to ensure that only packets containing video content with no unmet dependencies are transmitted. This three-pronged approach ensures that the streaming experiments adapt the SHVC streams efficiently under prevailing network conditions.

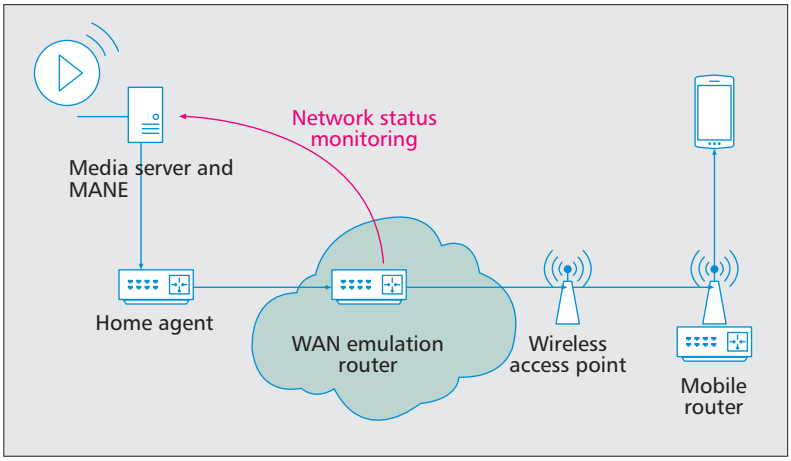


Figure 3. The wireless testbed used to compare SHVC and H.264/SVC stream adaptation. Stream adaptation takes place at the MANE, which is informed of the current network conditions by the WAN emulation router.

Sequence name	Spatial resolution	Frame rate
Traffic	2560 x 1600	30
PeopleonStreet	2560 x 1600	30
Cactus	1920 x 1080	50
BasketballDrive	1920 x 1080	50
Kimono	1920 x 1080	24

Table 1. Test sequences used for comparison.

COMPARING SHVC AND H.264/SVC

The following comparison of scalable video adaptation schemes employing H.264/SVC and SHVC highlights the bandwidth saving that could be achieved by adopting the newer SHVC encoding scheme. Video sequences drawn from the HEVC/SHVC standardization test sequences were encoded with both SHVC and H.264/SVC at two spatial resolutions chosen to be representative of popular display technologies. The full HD resolution (1920×1080) is widely available in a range of consumer devices including TVs, laptop computers, and high-end tablet/smartphone devices. A higher resolution (2560×1600) is currently popular on Apple retina displays and some other high-end portable devices.

The evaluation investigated spatial scalability and quality scalability, the two elements of SHVC that extend HEVC and have reference software implementations currently available. For spatial scalability, H.264/SVC sequences were encoded using a 2× configuration only. SHVC sequences were encoded using both 1.5× and 2× spatial scalability modes. Three QP testing points (22, 26, 30) were employed with both base layer and enhancement layer encoded at the same QP. The test sequences used are shown in Table 1.

The bandwidth requirements for each sequence encoded with 2× SHVC, 1.5× SHVC, and H.264/SVC spatial scalability are shown in Fig. 4 (for HD sequences) and Fig. 5 (2560×1600 sequences). We observed bandwidth savings in the range of 42 percent to 58 percent. On average, SHVC requires 47 percent less bandwidth than H.264/SVC. Spatial scalability in SHVC typically has a 12 percent lower bitrate than quality scalability. As with most video coding schemes, SHVC has different profiles designed for specific use cases. Although some variation in bandwidth requirement was seen between SHVC profiles, the differential with H.264/SVC remained consistent.

In both SHVC and H.264/SVC, quality scalability was evaluated at two testing points. In the first, the base layer was encoded with a QP of 32 and the enhancement layer with a QP of 30. The second testing point has a base layer QP of 22 and an enhancement layer QP of 20. Although employing a larger QP step leads to increased coding efficiency [11], this would be achieved at the expense of increased encoding time. When scalable streams are adapted at a MANE to reduce bandwidth requirements, employing spatial scalability results in a substantial reduction in bitrate of up to 63 percent, whereas using quality scalability, with a two QP differential, reduces the bitrate by a much lower margin of up to 28 percent. With respect to video quality, the objective PSNR metric was used to assess

The comparison between SHVC and H.264/SVC presented in this article has demonstrated the capabilities of an early implementation of SHVC adaptive streaming and highlighted its potential to improve streaming services in the consumer marketplace.

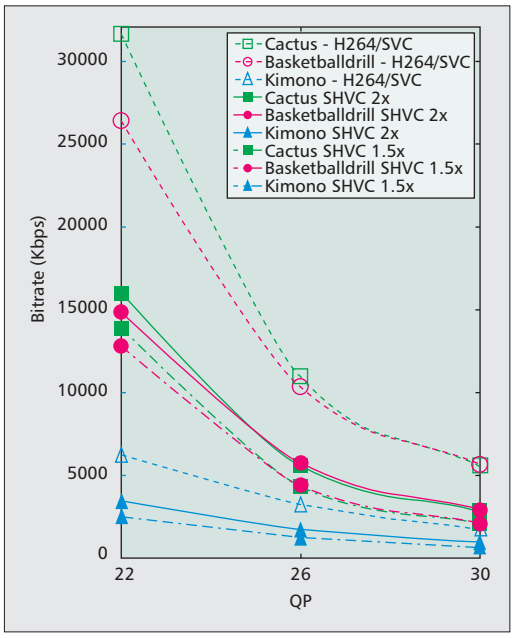


Figure 4. A comparison of bitrates for 1920 × 1080 sequences encoded using spatial scalability.

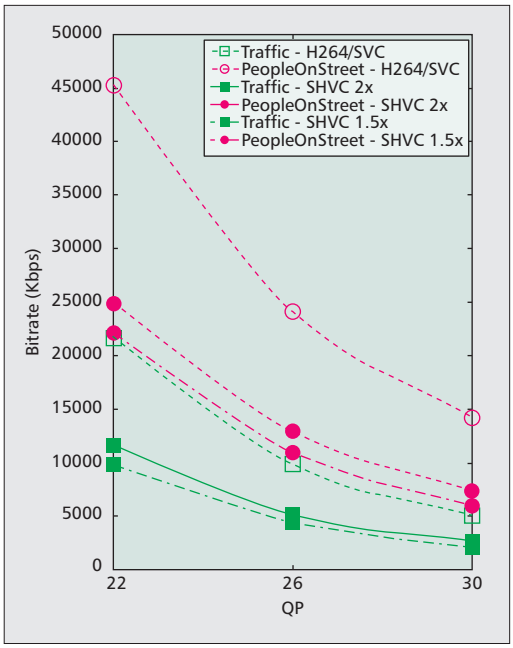


Figure 5. A comparison of bitrates for 2560×1600 sequences encoded using spatial scalability.

the impact of adaptation on scalable video streams, following the convention of JCT-VC in its comparative studies. Quality scalability produced better PSNR results with a modest degradation of 0.7dB of PSNR. Spatial scalability with a 2× configuration yielded an average PSNR reduction of 1.8 dB after adaptation. The drop in PSNR after adaption was not remarkably different between SHVC and H.264/SVC: on average H.264/SVC performed better than SHVC by a margin of 0.2 dB. However, when we applied a common bandwidth constraint of 3.3Mb/s to all of the encoded sequences, requiring that all but Kimono SHVC sequence be adapted by the MANE, the overall average quality of the H.264/SVC streams was 4.2dB lower than that of the SHVC streams. From these experiments, we can conclude that SHVC is capable of delivering high-quality video streams that need barely more than half the bandwidth of H.264/SVC with equal or better quality.

Opportunities and Challenges

The newly standardized H.265/HEVC and its proposed scalable extension will, when widely adopted, have a substantial impact on the design and development of the next-generation consumer products. The comparison between SHVC and H.264/SVC presented in this article has demonstrated the capabilities of an early implementation of SHVC adaptive streaming and highlighted its potential to improve streaming services in the consumer marketplace. Lower bandwidth requirements will mitigate the accelerating growth in video traffic and facilitate delivery of high-quality video over wireless channels. Higher spatial resolutions up to 8k will facilitate broadcast transmission of UHDTV content over terrestrial and satellite channels, and scalable video streams will enable transmission of the same content to many different consumer devices (from low-end phones to 8k televisions) over varied transmission networks. However, there are many challenges that will need to be addressed before SHVC-based streaming can be adopted as a core technology for consumer products. The first of these challenges arises from the computational complexity of HEVC and SHVC in comparison with previous encoding standards. Commercial broadcasters and content providers would need to embrace hardware-accelerated and/or distributed encoding architectures using parallel processing, perhaps cloud based, to facilitate encoding of UHDTV content and the additional representations required for scalable SHVC streaming or DASH. Real-time encoding and delivery of UHDTV content will pose particular challenges. Low-cost "on-chip" hardware decoders would be needed to drive the introduction of HEVC/ SHVC-enabled consumer products.

Although improved error resilience was one of the design objectives of the HEVC/SHVC standardization effort, considerable further work will be required to make robust, resilient delivery over impaired network channels a reality. In particular, robustness to packet loss in challenging wireless environments and novel methods of error concealment will be required.

Perhaps most interestingly of all, novel user-centric approaches to adaptive video streaming would need to be adopted. The current reliance on objective measurement of quality should be replaced by new metrics that fully consider the user's perception of not just visual quality, but his/her entire viewing experience. Current human vision system approaches to subjective or pseudo-subjective measurement are based on highly controlled laboratory experiments and largely ignore real-world factors. When considering immersive QoE, context factors (such as user

expectations of particular devices, service levels or contexts, for example, when travelling on public transport) that can detract from the user experience may also need to be explored. It may be time to consider moving to real-world 'total-experience' subjective evaluation of content delivery to everyday consumer devices in natural situations. By adopting user-centric approaches to service delivery, quality enhancements can be driven by user perception and expectation rather than the current quality of service mantras of availability, reliability, and integrity. Additionally, by employing a scalable video format it becomes practical to consider and measure the impact on a user's perception of quality when presented with different representations of the same video. The insights gleaned from such studies could be used, in conjunction with other factors such as video content type and prevailing network conditions, to develop new user-centered video streaming applications.

Acknowledgment

This work was partially funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J014729/1: Enabler for Next-Generation Mobile Video Applications and the UWS-BUU DBasS project.

References

[1] Cisco, "Visual Networking Index: Forecast and Methodology, 20122017," white paper, 2013.

[2] G. J. Sullivan *et al*., "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. Circuits Syst. Video Technol*., vol.22, no.12, Dec. 2012, pp. 1649–68.

[3] JCT-VC, "High Efficiency Video Coding (HEVC) Scalable Extension Draft 4," JCTVC-O1008_v3, 2013.

[4] ISO/IEC, Information Technology — Dynamic Adaptive Streaming over HTTP (DASH) — Part 1: Media Presentation Description and Segment Formats 23009-1, 2012.

[5] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, Sept. 2007, pp. 1103–20.

[6] JCT-VC, "Description of HEVC Scalable Extensions Core Experiment SCE1: Color Gamut and Bit-Depth Scalability," JCTVC-P1101_v2, 2014.

[7] J. Nightingale, Q. Wang, and C. Grecos, "SHVC Based Video Stream Adaptation in Hybrid Wired/Wireless Networks," *Proc. IEEE Int'l Symp. Personal Indoor and Mobile Radio (PIMRC) 2013*, 2013.

[8] I. Irondi, Q. Wang, and C. Grecos, "Empirical Evaluation of H.265/HEVC Based Real-Time Dynamic Adaptive Video Streaming over HTTP (HEVC-DASH)," *Proc. SPIE Photonics Europe 2014: Real-Time Image and Video Processing*, 2014.

[9] Minutes of VQEG Meeting Jan. 2014, available at http://www.its.bldrdoc.gov/vqeg/meetings/boulder,-co,-usa-january-21-25,-2014.aspx

[10] J. Nightingale, Q. Wang, and C. Grecos, "Optimized Transmission of H.264 Scalable Video Streams over Multiple Paths in Mobile Networks," *IEEE Trans. Consum. Electron*., vol. 56, no. 4, Nov. 2010, pp. 2161–69.

[11] B. Grkemli, Y. Adi, and A. M. Tekalp, "Effects of MGS Fragmentation, Slice Mode and Extraction Strategies on the Performance of SVC with Medium-Grained Scalability," *Proc. 17th IEEE Int'l Conf. Image Processing (ICIP)*, 2010, pp. 4201–04.

Biographies

James Nightingale (james.nightingale@uws.ac.uk) is a postdoctoral research fellow at the University of the West of Scotland (UWS), UK. His research interests include video streaming techniques, mobile networks, multihoming, and cloud computing. He is a member of IET and IEEE. He received his Ph.D. in video networking from UWS with the Outstanding Progression Award.

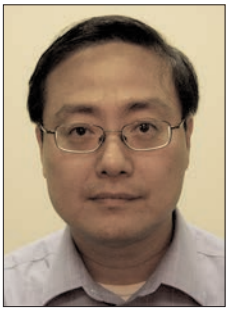
Qi Wang (qi.wang@uws.ac.uk) is a senior lecturer at the University of the West of Scotland (UWS), UK. He is the principal investigator of the UK EPSRC project "Enabler for Next-Generation Mobile Video Applications" (EP/J014729/1) and several industry-sponsored projects. He has over 70 publications on video networking, wireless/mobile networks, and cloud computing, which are his main research interests. He received his Ph.D. in mobile networking from the University of Plymouth, UK, with an ORS Award.

Christos Grecos (christos.grecos@uws.ac.uk) is a professor in visual communications standards, and the head of the School of Computing at the University of the West of Scotland (UWS), UK. His main areas of expertise are image and video compression algorithms and standards. He has over 150 publications, and has obtained significant research funding. His is a senior member of IEEE. He received his Ph.D. in image/video coding algorithms from the University of Glamorgan, UK.

Sergio Goma (sgoma@qti.qualcomm.com) is a senior director at Qualcomm Inc. in San Diego, USA, where he leads the multimedia R&D standardization group for imaging, driving Qualcomm's vision on future imaging technologies. His research interests include computational photography and programmable hardware architectures. Before joining Qualcomm he developed the image processing solution present in AMD's Imageon series of chips. He received a Ph.D. in reliability and fault tolerance of computers, and he holds several US patents on image processing algorithms and architectures.

By adopting user-centric approaches to service delivery, quality enhancements can be driven by user perception and expectation rather than the current quality of service mantras of availability, reliability and integrity.

AUTOMOTIVE NETWORKING AND APPLICATIONS


Wai Chen


Luca Delgrossi


Timo Kosch


Tadao Saito

In this 14th issue of the Automotive Networking and Applications Series, we are pleased to present three articles that address

- Release 1 of the cooperative intelligent transport systems (C-ITS) standards in Europe
- The implementation of a virtual traffic lights application with only partial market penetration
- The usage of multihop wireless communications for intra-car sensor networking

By timely information exchanges among vehicles, and between vehicles and roadway infrastructure, vehicles can transform from autonomous systems into cooperative systems, thereby enabling applications such as active road safety and traffic efficiency. Cooperative intelligent transport systems (C-ITS) standards are crucial to achieve interoperability among communications equipment made by different manufacturers for vehicles and roadway infrastructure. Release 1 of the C-ITS standards was completed in early 2014, and it covers base standards for ITS-G5 radio (also named wireless access in vehicular environment, or WAVE, in the United States), ad hoc networking and transport with GeoNetworking and Basic Transport Protocol (BTP), the facilities layer such as messaging protocols CAM and DENM, security, privacy and requirements for applications, among others. The first article, "Cooperative Intelligent Transport Systems Standards in Europe" by A. Festag, provides a comprehensive overview of release 1 of the C-ITS standards in Europe. The article first gives a brief overview of the C-ITS core standards set and compares it with the U.S. dedicated short-range communications (DSRC) standards in IEEE 1609 and SAE J2735. The author then provides more detailed overviews of the Release 1 C-ITS standards covering the access layer, the networking and transport layers, the facilities layer, and applications, security, and management in the subsequent sections of the article. The article concludes with a brief outlook on the expected C-ITS corridor pilots in Europe starting in 2015, as well as future standardization directions.

Virtual traffic lights (VTL) technology has recently been proposed as a new traffic management solution, where vehicles equipped with both the DSRC and VTL technologies will self-organize to dynamically elect a leader that serves as a virtual traffic light to decide the right of way at a given street intersection, thereby replacing the physical traffic lights. However, how to roll out the VTL technology without assuming 100 percent market penetration has been a main challenge. The second article, "Implementing Virtual Traffic Lights with Partial Penetration: A Road Map" by O. Tanguz *et al.*, shows that 100 percent market penetration is not necessary for deploying VTL technology. The authors first present a new method and system for the coexistence of VTL with the current infrastructure-based traffic lights system, starting with very small penetration ratios of VTL-equipped vehicles (i.e., vehicles with both DSRC and VTL technologies). The article then discusses simulation results to illustrate how the proposed game-theoretic VTL deployment methodology may work in various cases. The authors conclude with an extensive discussion on open challenges that need to be resolved in order to bring the VTL application into the marketplace.

Vehicles incorporate an increasing number of sensors to feed crucial situation data to electronic control units, whereas typically the sensors and control units are connected through physical wires, which increase the weight, maintenance, and cost of vehicles. The third article, "Intra-Car Multihop Wireless Sensor Networking: A Case Study" by M. Hashemi *et al.*, investigates the use of multihop wireless communication to support intra-car sensor networking and demonstrates the potential for significant reliability and energy consumption improvements over existing single-hop (star topology) approaches. The article first gives an overview of the existing intra-car wireless sensor network (WSN) studies and outlines a multihop intra-car WSN approach using the Collection Tree Protocol. The authors then show, via extensive test results of two in-car networks (a suspension network and an engine network) under static and dynamic vehicular conditions, that the multihop WSN approach can achieve higher packet deliv-

ery rates across all sensor nodes as well as lower radio energy consumption than that of the star- topology (single-hop) approaches.

We thank all contributors who submitted manuscripts for this Series, as well as all the reviewers who helped with thoughtful and timely reviews. We thank Dr. Sean Moore, Editor-in-Chief, for his support, guidance, and suggestions throughout the process of putting together this issue. We also thank the IEEE publication staff, particularly Ms. Jennifer Porcello, for their assistance and diligence in preparing the issue for publication.

BIOGRAPHIES

WAI CHEN (waichen@ieee.org) received his B.S. degree from Zhejiang University, and M.S., M.Phil., and Ph.D. degrees from Columbia University, New York. He is chief scientist of China Mobile Research and general manager of China Mobile Internet-of-Things Research Institute. Previously, he was vice president and group director of ASTRI, Hong Kong, and chief scientist and director at Telcordia (formerly known as Bellcore), New Jersey. While at Telcordia, he led a vehicular communications research program for over 10 years in collaboration with a major automaker on automotive networking technologies for vehicle safety and information applications. He was Principal Investigator of several government funded projects on advanced networking technologies research. He was General Co-Chair for the IEEE Vehicular Networking Conference (IEEE VNC 2009–2013) and a Guest Editor for a Special Issue on Vehicular Communications and Networks of the *IEEE Journal on Selected Areas in Communications* (2011). He also served as a Guest Editor for a Special Issue on Inter Vehicular Communication (IVC) for *IEEE Wireless Communications* (2006). He was an IEEE Distinguished Lecturer (2004–2006), Co-Chair of the Vehicle-to-Vehicle Communications Workshop (IEEE V2VCOM 2005–2008) co-located with the IEEE Intelligent Vehicles Symposium, Co-Chair of the IEEE Workshop on Automotive Networking and Applications (IEEE AutoNet 2006–2008) co-located with IEEE GLOBECOM, and Vice Chair of the Technical Program Committee for Vehicular Communications of the IEEE Vehicular Technology Conference (IEEE VTC-Spring 2009).

LUCA DELGROSSI is manager of the Vehicle-Centric Communications Group at Mercedes-Benz Research & Development North America Inc., Palo Alto, California. He started as a researcher at the International Computer Science Institute (ICSI) of the University of California at Berkeley and received his Ph.D. in computer science from the Technical University of Berlin, Germany. He served for many years as a professor and associate director of the Centre for Research on the Applications of Telematics to Organizations and Society (CRATOS) of the Catholic University at Milan, Italy, where he helped create and manage the Master's in Network Economy (MiNE) program. In the area of vehicle safety communications, he coordinated the Dedicated Short Range Communications (DSRC) Radio and On-Board Equipment work orders to produce the DSRC specifications and build the first prototype DSRC equipment as part of the Vehicle Infrastructure Integration (VII) initiative of the U.S. Department of Transportation. The Mercedes-Benz team in Palo Alto is a recognized leader in the R&D of vehicle-to-infrastructure as well as vehicle-to-vehicle communications safety systems.

TIMO KOSCH works as a team manager for BMW Group Research and Technology, where he is responsible for projects on distributed information systems, including such topics as cooperative systems for active safety and automotive IT security. He has been active in a number of national and international research programs, and serves as Coordinator for the European project COMeSafety, co-financed by the European Commission. He is also currently heading the system development for a large German Car2X field test. For more than three years, until recently, he chaired the working group Architecture and was a member of the Technical Committee of the Car-to-Car Communication Consortium. He studied computer science and economics at Darmstadt University of Technology and the University of British Columbia in Vancouver with scholarships from the German National Merit Foundation and the German Academic Exchange Service. He received his Ph.D. from the computer science faculty of the Munich University of Technology.

TADAO SAITO [LF] received a Ph. D degree in electronics from the University of Tokyo in 1968. Since then he has been a lecturer, an associate professor, and a professor at the University of Tokyo, where he is now a professor emeritus. Since April 2001 he is chief scientist and CTO of Toyota InfoTechnology Center, where he studies future ubiquitous information services around automobiles. He has worked in a variety of subjects related to digital communication and computer networks. His research includes a variety of communication networks and their social applications such as ITS. Included in his past study, in the 1970s he was a member of the design group of the Tokyo Metropolitan Area Traffic Signal Control System designed to control 7000 intersections under the Tokyo Police Authority. Now he is Chairman of the Ubiquitous Networking Forum of Japan working on a future vision of the information society. He is also Chairman of the Next Generation IP Network Promotion Forum of Japan. He wrote two books on electronic circuitry, four books on computers, and two books on digital communication and multimedia. From 1998 to 2002 he was chairman of the Telecommunication Business Committee of the Telecommunication Council of the Japanese government and contributed to regulatory policy of telecommunication business for broadband network deployment in Japan. He is also the Japanese representative to the International Federation of Information Processing General Assembly and Technical Committee 6 (Communication System). He is an honorary member and Fellow of IEICE of Japan.

Cooperative Intelligent Transport Systems Standards in Europe

Andreas Festag

ABSTRACT

Information exchange among vehicles, and between vehicles and the roadside infrastructure is commonly regarded as a base technology to sustainably reduce road accidents and improve traffic efficiency. After more than a decade of research and development efforts, a technological basis has been established that applies WiFi-based, wireless communication in the 5.9 GHz frequency band, ad hoc communication and dedicated message sets, as well as management and security procedures. In Europe, Release 1 of standards for cooperative systems has been completed, indicating deployment of a basic system starting in 2015. This article provides a comprehensive overview of standards and complementary industry specifications for cooperative systems in Europe, covering relevant aspects of access technologies, network and transport protocols, facilities, applications, security, and management.

INTRODUCTION

Vehicles are getting safer, cleaner, and more intelligent. Various sensors and assistant systems enable vehicles to monitor their environment. By means of information exchange among vehicles, as well as between vehicles and the roadside infrastructure, vehicles transform from autonomous systems into cooperative systems. Inter-vehicle communication is a cornerstone of intelligent transportation Ssystems (ITS), commonly referred to as *cooperative ITS* (*C-ITS*) or *car-2-X communication*.

The development of C-ITS is primarily driven by applications for active road safety and traffic efficiency, which help drivers to be aware of other vehicles, disseminate warnings about road hazards, and provide real-time information about traffic conditions for speed management and navigation. Typically, these C-ITS applications rely on always-on connectivity among the vehicles in the vicinity, including the roadside infrastructure, and frequent data exchange. Additionally, Internet access and location-based services, such as for point-of-interest notification, road access control, and parking management, improve the driving convenience. Among the various possible communication technologies for ITS, a dedicated variant of IEEE 802.11, an allocated frequency band at 5.9 GHz for road safety and traffic efficiency applications, ad hoc networking, and C-ITS specific message sets have emerged as the current state of the art.

In Europe, the first research programs for cooperative ITS date back to the 1980s; the European project PROMETHEUS (1987–1994) marked the beginning of a cooperative driving system using inter-vehicle communication in the 57 GHz frequency band. By 2000, a new wave of research and development activities in academia and industry was initiated worldwide, triggered by the availability of GPS, embedded systems, and WiFi. In Europe, more than 40 different projects on C-ITS have been initiated since 2000. Starting with initial feasibility studies, such as FleetNet and NoW, projects greatly contributed to the current technology state and standardization, for example, SAFESPOT, GeoNet, SEVECOM, CoVeL, and COMeSafety. Finally, field operation tests (DRIVE C2X, SIM-TD, SCORE@F, etc.) validated and assessed the potential positive impact of C-ITS on safety and traffic efficiency at various test sites across Europe. Further projects have been initiated to study cooperative automated driving, such as the AutoNet2030 project.

C-ITS standards are essential to achieve interoperability among communication devices from different manufacturers for vehicles and roadside infrastructure. In Europe, standards are being developed by the European standardization organizations (ESOs) European Telecommunications Standards Institute (ETSI) and omité Européen de Normalisation (CEN) in their respective technical committees (TCs) ETSI TC ITS and CEN TC 278, *Road Transport and Telematics*, the latter in close liaison with International Organization for Standardization (ISO) TC 204. The standardization scope covers all types of transport, including rail, water, and air transport (ETSI) as well as tolling systems and road infrastructure (CEN); nevertheless, the focus in the past was clearly on cooperative road vehicles. ESOs produce standards of different types, from which the European Norm (EN) are approved by the national standardization organizations (NSOs) of the EU member and associated states and made legally binding. In 2010 the European Commission issued a mandate to the

The author is with the Technical University of Dresden and NEC Laboratories Europe.

This work was also supported by the European Commission under AutoNet2030 – Co-operative Systems in Support of Networked Automated Driving by 2030, a collaborative project part of the Seventh Framework Programme for research, technological development and demonstration (Grant Agreement NO. 610542, URL http://www.autonet2030.eu). The author would like to thank all partners within AutoNet2030 for their cooperation.

0163-6804/14/$25.00 © 2014 IEEE

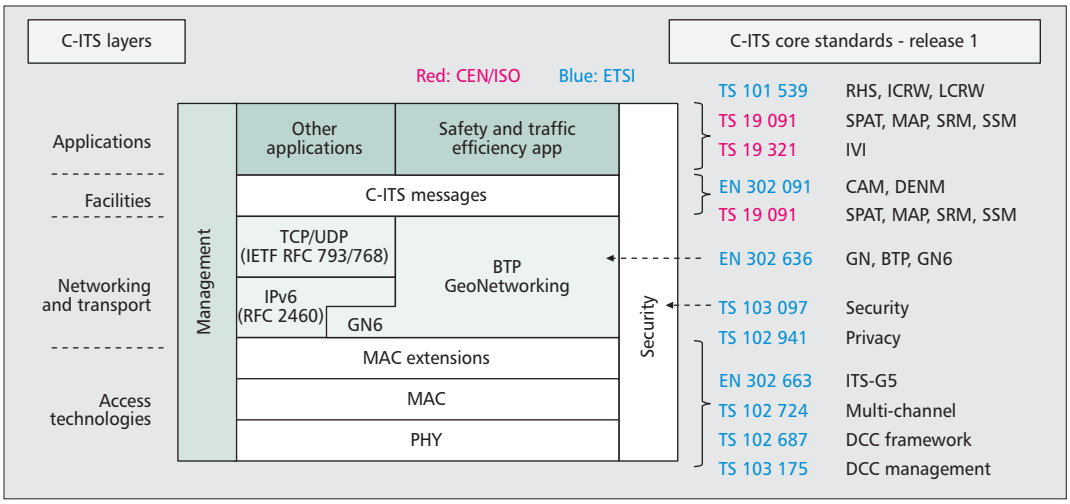


Figure 1. Protocol stack and Release 1 core standards for C-ITS in Europe.

Typically, the GeoNetworking stack is used for ad hoc communication over ITS-G5 utilizing the geo-addressing, and IPv6 for communication with an IP-based infrastructure over cellular networks. IPv6 packets can also be transmitted over GeoNetworking, for which the adaptation sublayer GN6 has been designed.

ESOs [1] for the development of a minimum and consistent set of standards for C-ITS. The mandate implied a common basis for national standardization in Europe and therefore prevented conflicting national standards. It was completed in 2013 with the announcement of Release 1 of the standards [2].

In Europe, the standardization efforts are driven by the European *Car-2-Car Communication Consortium* (*C2C-CC*) [3], an industry consortium of automobile manufacturers, suppliers and research organizations, *ERTICO*, an European organization of stakeholders with public and private partners, as well as by ETSI's Center for Testing and Interoperability, *ETSI CTI*. In 2013, automobile manufacturers in C2C-CC signed an agreement for the introduction of the system in Europe starting in 2015. Deployment plans are being developed in the Amsterdam Group [4], a strategic alliance of stakeholders of C-ITS in Europe, *CEDR*, *ASECAP*, *POLIS*, and *C2C-CC*.

This article gives a comprehensive overview of Release 1 C-ITS standards in Europe and their profiling by industry consortia for initial deployment by 2015. The second section gives an overview about the standards set and briefly compares it with the IEEE 1609 standard family. The following sections provide details about standards for access layer 3, networking and transport, facilities, applications, and security and management. The final section concludes and provides an outlook on deployment and standardization beyond Release 1.

Overview of C-ITS Standards

The C-ITS standards follow a general architecture, specified in ETSI EN 302 665 and ISO 21217, with the *ITS station* as the core element, representing vehicle, personal (mobile personal devices), roadside (infrastructure), and central (back-end systems and traffic management centers) subsystems [5]. For C-ITS, the ISO OSI reference model was adapted to cover horizontal layers for access technologies, networking and transport, facilities and applications, and vertical entities for management and security.

Figure 1 illustrates the protocol stack for vehicle and roadside ITS stations, and lists the Release 1 core standards with their shorthand names for the European C-ITS Release 1. The access technologies layer primarily utilizes a specific set of options of the IEEE 802.11 standard, that is, ITS-G5 (where G5 stands for the 5 GHz frequency band). In the United States, this set is named Wireless Access in Vehicular Environment (WAVE), formerly referred to as the IEEE 802.11p amendment and now integrated into the IEEE 802.11-2012 standard release. The European variant, ITS-G5, is derived from WAVE and adapted to European requirements. Other access technologies, such as cellular networks, are not excluded, but are out of the scope of this article. The networking and transport layer has two columns: GeoNetworking and Basic Transport Protocol (BTP). The other column employs the Internet protocols, in particular IPv6 with UDP, TCP, or potentially other transport protocols such as SCTP, and IP mobility extensions (Mobile IPv6 and its extensions for network mobility, NEMO). The choice of the communication profile, whether GeoNetworking or IPv6, lies in the application. Typically, the GeoNetworking stack is used for ad hoc communication over ITS-G5 utilizing geo-addressing, and IPv6 for communication with an IP-based infrastructure over cellular networks. IPv6 packets can also be transmitted over GeoNetworking, for which the adaptation sublayer GN6 has been designed.

On top of the network and transport layer, standards for facilities layer protocols enable application functionality. The CAM protocol conveys critical vehicle state information in support of safety and traffic efficiency application, with which receiving vehicles can track other vehicles' positions and movement. The DENM protocol disseminates event-driven safety information in a geographical region. Further message types are being standardized for vehicle-to-infrastructure communication. Applications are not fully standardized; instead, standards specify the minimum requirements for three groups of applications: *road hazard signaling* (*RHS*) comprises 10 different use

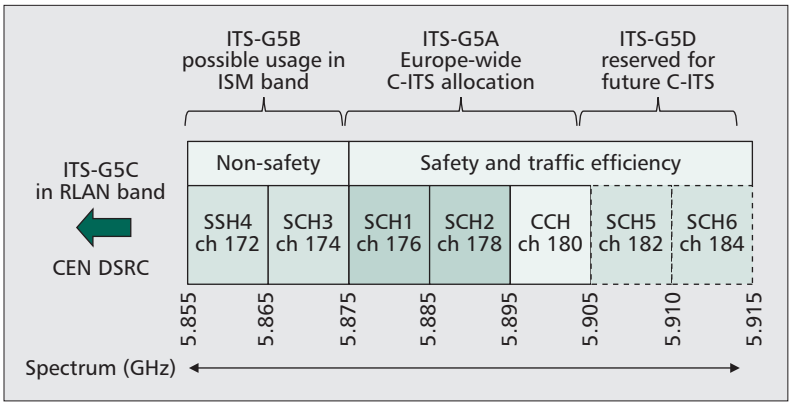


Figure 2. European frequency allocation for road safety and traffic efficiency (ETSI EN 302 571).

cases for initial deployment, including emergency vehicle approaching, hazardous location, and emergency electronic brake lights. The other two groups, *intersection collision risk warning* (ICRW) and *longitudinal collision risk warning* (LCRW), refer to potential vehicle collisions at intersections and rear-end/head-on collisions. Security- and privacy-related standards enable asymmetric cryptography and changing pseudonyms. Management standards mainly cover support for decentralized congestion control and communication profile management. A series of test standards provide specifications to verify the conformance of an implementation to the base standards and enable plug-tests for the testing of interoperability among implementations from different vendors. Further industry specifications for profiling and missing standards complete the set.

A comparison of the European C-ITS with the U.S. dedicated short range communication (DSRC) standards in IEEE 1609 and SAE J2735 [6] reveals many similarities. Both approaches operate in the 5.9 GHz frequency band with several 10 MHz channels and rely on the OCB mode of IEEE 802.11, whereas the European variant *ITS-G5* takes into account specific requirements for Europe and also incorporates service channels at the RLAN 5.4–5.7 GHz band for ITS applications. Also, the security and privacy approach is similar. Standards for higher protocol layers are different: the U.S. IEEE 1609 specifies a broadcast protocol for ad hoc routing optimized for short packet headers, called *Wave Short Message Protocol* (WSMP). The *ETSI GeoNetworking* standards specify an ad hoc routing protocol for single- and multihop communication with geographical addressing. Furthermore, the U.S. approach largely relies on the *basic safety message* (BSM) for collision avoidance applications. In contrast, C-ITS uses several safety message types including CAM for periodic and DENM for event-driven safety information. Both approaches are the subject of harmonization efforts at the governmental level between the United States and Europe, also including Japan, and at the industry level between C2C-CC and CAMP on the U.S. side. A major achievement is the alignment of the C-ITS Common Data Dictionary (CDD) in ETSI, which specifies the data elements for the CAM, DENM, and other messages, with the SAE 2735 message set. Overall, the similarities between the European C-ITS and U.S. DSRC standards prevail and enable multi-mode or dual stack implementations at reasonable costs, although major conceptual differences, particularly in information dissemination, still remain.

Access Layer Standards

Three frequency bands in the 5 GHz band were allocated for ITS in Europe in 2008 (Fig. 2), aligned with similar efforts in North America and Japan: *ITS-G5A* has 30 MHz with 10 MHz channel spacing. The upper channel in ITS-G5A is named the *control channel* (*CCH*) and used as the primary safety channel. The others are additional service channels. *ITS-G5B* spans 20 MHz with two service channels of 10 MHz each and is dedicated to non-safety C-ITS applications. The 255 MHz wide band *ITS-G5C* is shared with the radio local area network (RLAN) band used by WiFi devices and can have 10 or 20 MHz channels. In ITS-G5C, devices must adhere to the dynamic frequency selection (DFS) method, well known for WiFi devices, which protects radar systems operating in the same band. As this method requires a DFS master, the usage of ITS-G5C is practically restricted to vehicle-to-infrastructure communication with a C-ITS roadside unit as a DFS master. However, up to now, the effectiveness of DFS with highly mobile devices in vehicles is not clear. Two more channels (*ITS-G5D*) are foreseen for future C-ITS systems.

In Europe, the usage of the allocated bands is regulated by harmonized standards, a specific form of European norms, which ensure the compliance of radio equipment with legislative directives. The harmonized standard for ITS-G5A EN 302 571 allows for up to 23 dBm/MHz transmission power and limits the emission to adjacent bands accordingly [7]. The spectrum mask is limited to –65 dBm in the 5.795–5.805 GHz band, in which CEN DSRC, the European tolling system, operates. The maximum transmit power is further restricted per service channel in order to protect the important CCH and restrict the out-of-band leakage to the CEN DSRC band: on the CCH and SCH1 a transmit power of 23 dBm/MHz is allowed, on the SCH2 and SCH3 only 13 dBm/MHz. Requests to spectrum regulators for extension of WiFi operation to the 5 GHz band for high data rate communication have triggered discussions about coexistence of C-ITS and WiFi in the same band. Although studies are still ongoing, effective protection of the ITS-G5 bands with low latency requirements is technically challenging.

The physical transmission in ITS-G5 is derived from IEEE 802.11a. It uses orthogonal frequency-division multiplexing (OFDM) with 52 subcarriers, of which 48 are for data and 4 for pilots. Compared to the typical 20 MHz channels in IEEE 802.11, with 10 MHz channels the subcarrier spacing is halved, and the timing parameters are doubled, yielding an OFDM symbol duration of 8 μs including a cyclic prefix of 1.6 μs, which is aligned with the expected delay spread at a communication range of less than 1

km. The symbol duration in the 10 MHz channel configuration is also small enough to cope with inter-carrier interference caused by Doppler spread, that is, the coherence time is $T_C = f_d T_{OFDM} \approx 0.0134 \ll 1$ for a relative speed of 260 km/h between two vehicles. At a default data rate of 6 Mb/s the coherence time corresponds to roughly 400 bytes, which may result in variations of the channel estimation during the reception of a large frame in high-speed scenarios [8].

ITS-G5 applies a basic ad hoc mode, which is referred to *outside the context of a BSS* (*OCB*) in standards. The OCB mode simplifies operation compared to a wireless network known as the *basic service set* (*BSS*) in IEEE 802.11 terminology, disables management procedures, such as channel scanning, authentication, and association, and uses a wildcard BSS identifier. OCB-enabled ITS-G5 stations can transmit messages directly and immediately without time-consuming delays for the exchange of control frames.

For medium access, ITS-G5 introduces the same scheme as specified in IEEE 802.11, that is, carrier sense multiple access with collision avoidance (CSMA/CA), with one medium access control (MAC) entity per channel. The scheme is extended by quality of service (QoS) support from IEEE 802.11, known as *enhanced distributed channel access* (*EDCA*), which provides different priorities for channel access with specific parameters for contention window size and idle time (inter-frame spaces) per priority class. The ITS-G5 frames are assigned to EDCA queues based on the traffic class (TC) parameter chosen by the facilities layer with TC values for CAMs, high- and low-priority DENMs, and so on.

On top of the medium access entity with the EDCA queues, the MAC layer extensions for ITS-G5 provide two main functions: *gatekeeper* and *multi-channel operation* (*MCO*). The gatekeeper ensures that upper layer entities transmit packets within a maximum rate bound for a TC; it is essentially a set of on-off queues on top of the EDCA queues. MCO controls an ITS-G5 transceiver to tune to a specific service channel in a dual-transceiver ITS-G5 configuration, where the first transceiver is fixed to the CCH, and the other transceiver may dynamically switch among service channels. The standards for both functions, gatekeeper (ETSI TS 102 687) and MCO (ETSI TS 102 724), are currently under revision.

ITS-G5 uses the standard IEEE 802.2 protocol supplemented by the *Subnetwork Access Protocol* (SNAP). GeoNetworking uses LLC unacknowledged connection (Type 1) service with unnumbered information (UI) frames and Ethertype 0x8947.

Networking and Transport Layer Standards

The networking and transport layer standards belong to the standard series EN 302 636 and cover requirements (part 1), scenarios (part 2), and the overall networking architecture (part 3). Part 4 specifies the networking protocol and is separated into media-independent (subpart 4-1) and media-dependent operations for ITS-G5 (subpart 4-2). Although split up in parts, both build a single protocol entity. The split allows for future media-specific extensions over wireless media other than ITS-G5. The transport layer standards for BTP and GN6 are specified in parts 5 and 6 of the series, respectively.

GeoNetworking is a routing protocol that provides packet delivery in an ad hoc network without a coordinating infrastructure. It utilizes geographical positions for addressing and forwarding. The addressing capabilities facilitate sending a packet to an individual ITS station with its geographical position or to a geographical target area described by geometric shapes (circle, rectangle, ellipse; see ETSI EN 302 931); the latter implies both broadcast and anycast to nodes inside the target area (area forwarding mode), as well as support for the transport of packets toward the area if the source is located outside (line forwarding mode). Altogether, GeoNetworking supports five packet handling modes: geo-unicast, geo-broadcast, geo-anycast, single-hop broadcast, and topologically-scoped broadcast, the latter two not having the geographical addressing. Geo-broadcast packets are used to distribute event-driven messages of type DENM, and periodically triggered CAMs are carried by single-hop broadcast packets.

GeoNetworking enables forwarding of packets *on the fly* without the need to establish and maintain routes. The GeoNetworking standard EN 302 636-4-1 specifies several forwarding algorithms with increasing protocol functionalities and efficiency. For geo-broadcast packets, three algorithms are specified: *simple geo-broadcast* applies a flooding scheme that restricts rebroadcasting by the geographical borders of the target area, and duplicate packet detection based on source ID and packet sequence numbers. With *contention-based forwarding* (CBF) a node broadcasts the packet to all neighbors, which buffer the packet and contend for packet forwarding: Each candidate forwarder starts a timer that is inversely proportional to its forwarding progress (i.e., the distance between the local and previous sender positions). The node with the shortest timer wins the contention and rebroadcasts the packet. When the other contending nodes overhear a forwarded packet, they stop the timer and remove the packet from their buffer. The *advanced forwarding* algorithm combines CBF with a sender-based selection of the next hop, where the neighbor with the most progress is chosen (also referred to as greedy forwarding, GF). The advanced forwarding algorithm is illustrated in Fig. 3 in an example scenario. The source node S detects a safety event, creates a geo-broadcast packet, and selects F1 as the next hop using the GF algorithm. F1 then forwards the packet without buffering. The other nodes inside a defined sector of the source's direct communication range process and buffer the packet, and forward it if their position-dependent CBF timer expires. The advanced forwarding also allows for redundant retransmission by several different nodes up to a configurable threshold. The redundant transmission improves the reliability of geo-broadcast packet dissemination and controls the number of

Altogether, GeoNetworking supports five packet handling modes: geo-unicast, geo-broadcast, geo-anycast, single-hop broadcast, and topologically-scoped broadcast, the latter two not having the geographical addressing.

retransmissions, avoiding well-known broadcast storms. For geo-unicast, corresponding forwarding algorithms, that is, *GF* and *CBF*, are defined.

A GeoNetworking packet is composed of three headers; basic, common and extended. The basic and common headers carry fields that are needed by all packet types. The extended header is specific for geo-unicast, geo-broadcast, and so on, and covers, for example, fields for the definition of the geo-area. The separation of basic and common headers has security reasons: a digital signature and certificate is generated by the packet source over the common and extended headers (and payload), such that fields in the basic header can be modified by a forwarder (e.g., the hop-count value can be decreased by every forwarder without the need to regenerate the signature [9]).

On top of the GeoNetworking protocol, BTP (EN302 636-5-1) multiplexes/demultiplexes facility-layer messages and provides a connectionless, unreliable end-to-end packet transport similar to UDP. It adopts the concept of ports from the IP suite and assigns well-known ports for the relevant facility-layer message types. Alternatively to BTP, GN6 (EN 302 636-6-1) enables sub-IP multihop delivery of IPv6 packets without modifications of IPv6. It adapts the stateless address auto-configuration known from IPv6 and extends the concept of an IPv6 link to geographical areas that are associated with an IPv6 point of attachment. GN6 introduces an adaptation sublayer, referred to GN6ASL that presents a flat network topology to IP [10, 11].

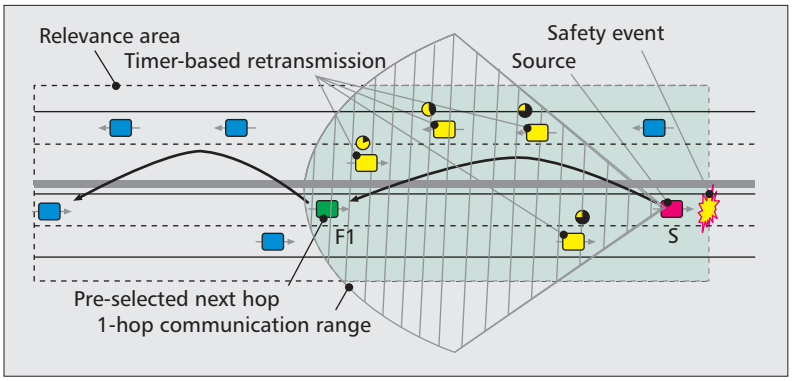


Figure 3. Advanced forwarding algorithm for geo-broadcast (ETSI EN 302 636-4-1).

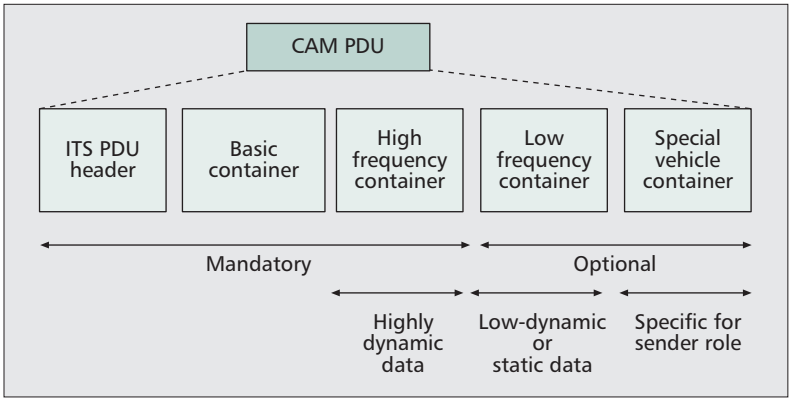


Figure 4. CAM structure (ETSI EN 302 637-2).

Facilities Layer Standards

The facilities layer standards specify requirements and functions supporting applications, communication, and information maintenance. The standards cover messaging protocols, position and time management, location referencing, sensor data fusion in a *local dynamic map* (LDM), and others. The most relevant standards are those for the C-ITS messaging, which are presented below.

CAM is a periodic message that provides status information to neighboring ITS stations. Its transmission is activated when a vehicle is in a safety-relevant context (basically, when the engine is running). A CAM is composed of an ITS PDU header and several containers (Fig. 4) that group the data fields by the role of the sender and frequency of their appearance in the message. The ITS PDU header carries protocol version, message type, and sender address; the basic container has station type and its position. In order to reduce the size of the CAM, the high-frequency container carries mainly highly dynamic data (e.g., vehicle heading, speed, and acceleration) and is sent in every CAM. The low-frequency container has data with less safety relevance (e.g., vehicle role) or may have a large size (e.g., path history) and is therefore not always added to the CAM, but sent. The special vehicle containers are optionally added if needed for the sender's role, such as for public transport, dangerous goods, road works, or rescue. The container concept ensures a flexible message format that can be adapted to the needs of the sending and receiving vehicle, while minimizing the load on the wireless channel.

The CAM rate is determined by CAM generation rules and can vary between the lower and upper limit of the CAM period T_{Min} = 100 ms and T_{Max} = 1 s (corresponds to a CAM rate of 1 to 10 in 1 s), controlled by the vehicle dynamics, application, and congestion status of the wireless channel. As illustrated in Fig. 5, the conditions are sampled at small intervals (minimum 10 Hz). If the vehicle dynamics exceed the predefined thresholds for heading, movement, and acceleration, a CAM is generated. The minimum and maximum time period between two CAMs can then further be restricted by the needs of DCC and the applications to T_{DCC} and T_{APP}, respectively: If the load on the wireless channel is high, the minimum time period is increased, whereas the application is able to decrease the maximum time period if required by the safety situation. A low-frequency container and special vehicle container are included if at least 500 ms has passed since the last CAM generation.

DENM is an application-controlled, safety event-triggered message. When a vehicle detects a safety situation, the DENM protocol assigns an *action identifier* that is unique for the detecting ITS station. Unlike the CAM broadcast over a single ITS-G5 hop, the DENM gets assigned a relevance area for dissemination and can be transported over several wireless ITS-G5 hops, typically through the geo-broadcast mode of the GeoNetworking protocol. Similar to the CAM, the DENM is organized in containers with a prepended ITS PDU header (Fig. 6). The man-

agement container — with fields for action identifier, detection time, event position, and so on — is mandatory, all other containers are optionally added if needed by the application. The situation container has fields to describe the event by a predefined code for the causing event as well as related events (e.g., linked events or an event history). The location container carries fields for the event speed, heading, and traces. An a la carte container can be added to transmit application-specific contents, such as for lane position, impact reduction, and road works, among others.

The DENM protocol can handle an event life cycle: an event with a specific action ID can be triggered and then updated by the originator of the DENM; the event updates are distinguished by an increasing value of a data version field. An event can also be canceled by the originator or negated by a third ITS station.

The DENM protocol specification has several mechanisms for information dissemination to keep the safety information in the relevant area during the event lifetime. The originator can repeat a DENM, typically at a lower frequency than a CAM, to ensure that vehicles entering the relevant area later can receive the information. Optionally, another ITS station than the originator can overtake the repetition of the DENM message in case the originator fails to repeat the DENM (e.g., if it is broken or has left the relevant area).

In addition to CAM and DENM standards, further messages are being standardized in CEN TC 278/ISO TC 204 for static road topology data (MAP), dynamic traffic light data (signal phase and timing, SPAT), priority and preempted access of special vehicles (SRM, SSM), probe vehicle data (PVD, PDM), and in-vehicle information (IVI).

Applications, Security, and Management Standards

For the initial deployment of C-ITS, a basic set of applications (BSA) has been identified (ETSI TR 101 638) and classified into four application groups: active road safety, cooperative traffic efficiency, cooperative local services, and global Internet services. Applications are not fully standardized; instead, the standards specify the minimum requirements for three groups of applications: RHS, ICRW, and LCRW. Road hazard signaling (RHS) comprises 10 different use cases that are relevant for initial deployment, including emergency vehicle approaching, hazardous location, and emergency electronic brake lights. The other two groups, intersection collision risk warning (ICRW) and longitudinal collision risk warning (LCRW), refer to potential vehicle collisions at intersections and rear-end/head-on collisions, respectively. In addition to the requirements from SDOs, the C2C-CC has defined triggering conditions for its day 1 applications that specify the behavior of use cases for the sender, including pre-conditions, process flow, message parameters, and information quality requirements for the specific use case.

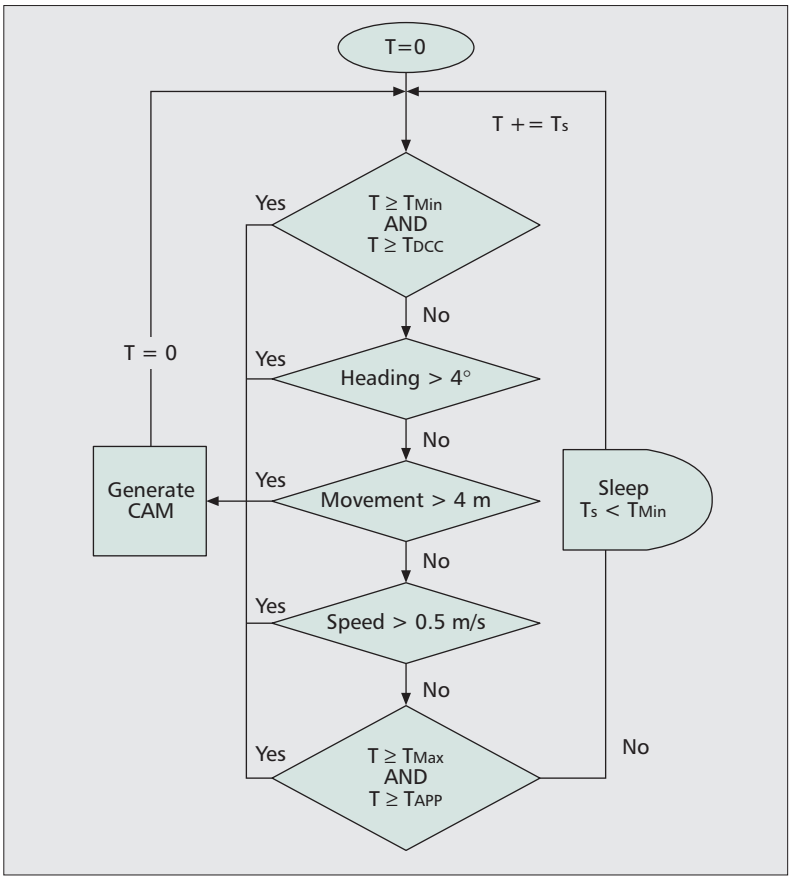


Figure 5. CAM generation rules (ETSI EN 302 637-2).

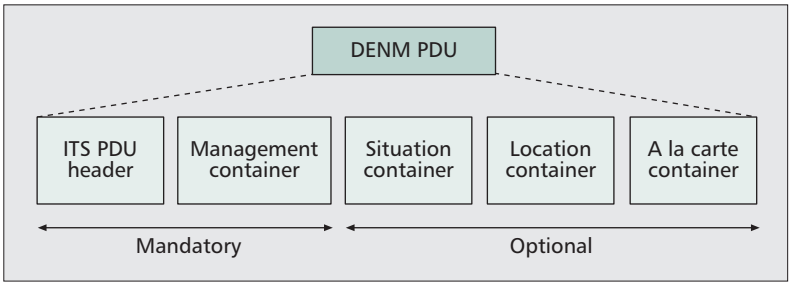


Figure 6. DENM structure (ETSI EN 302 63-3).

C-ITS standards specify mechanisms for security and privacy protection [12]. Based on the security architecture in ETSI TS 102 940, ETSI TS 102 097 specifies private key infrastructure (PKI) enrollment and authorization management protocols, ETSI TS 102 941 confidentiality, and ETSI TS 102 942 data integrity. The core security standard is ETSI TS 103 097 for the security header and certificate format for asymmetric cryptography with elliptic curves. The standards are complemented by C2C-CC specifications for PKI, TAL, and PP. The PKI specifications define the protocols among the certificate authorities and with the ITS stations, including root, long-term, and pseudonym certificate authorities. The trust assurance levels (TALs) define security levels of an in-vehicle C-ITS system, from a basic TAL protecting the software up to a very high TAL with a tamper-resistant hardware model, and shielding the

For DCC management, ETSI TS 102 687 defines a toolbox-like framework to control the channel load by transmit power, message rate, and other parameters. The standard introduces a state machine for DCC with relaxed, active, and restrictive states depending on the actual channel busy time.

involved in-vehicle sensors and control units. The protection profile (PP) then comprises all measures for security and privacy of a given TAL.

For DCC management, ETSI TS 102 687 defines a toolbox-like framework to control the channel load by transmit power, message rate, and other parameters. The standard introduces a state machine for DCC with relaxed, active, and restrictive states depending on the actual channel busy time. Every channel state enforces a predefined set of parameters for the upper or lower threshold of DCC parameters, but does not specify the exact DCC algorithm. ETSI TS 103 175 defines an ITS station-internal management entity that evaluates the congestion status for the ITS-G5 channels based on information from different layers. The DCC-related specifications are currently being revised. It is expected that message rate control is being applied as the main DCC mechanism together with additional mechanisms to ensure coexistence of C-ITS and the CEN DSRC road tolling system that operates in the adjacent 5.8 GHz frequency band. As an alternative to the state-based DCC approach, algorithms with linear control of parameters are considered, in particular LIMERIC [13] and PULSAR [14].

Conclusions and Outlook

C-ITS has developed into a mature technology that can enable a wide range of innovative applications. To achieve interoperability, standards are essential. Release 1 of C-ITS standards was completed in early 2014, covering base standards for ITS-G5 radio, ad hoc networking and transport with GeoNetworking and BTP, facilities layer standards, in particular the messaging protocols CAM and DENM, security, privacy, and requirements for applications. The base standards are complemented by a set of test specifications. The maturity of the standards has been validated by conformance tests and plug-tests for interoperability with prototype implementations from different vendors. Field operational tests across Europe (DRIVE C2X, SIM-TD, and SCORE@F) have implemented and validated the standards in large-scale studies to assess the impact of C-ITS on safety and traffic efficiency.

Based on Release 1 of C-ITS standards, the C2C-CC has derived a profile that restricts the large list of standards and parameters to a practical set, and complementary missing specifications for security, management, and applications. It is expected that the deployment of the C-ITS profile will start in Europe in 2015. To ease the system introduction, corridor pilots are planned that will provide C-ITS services on major European highways. The forerunner is the trilateral C-ITS corridor interconnecting Vienna–Frankfurt–Rotterdam starting in 2015. Further corridor pilots are planned in France, Sweden, and other countries, as well as in selected cities. It is important to note that the developed C-ITS message sets are media-agnostic and expected to be reused for media other than ITS-G5, in particular 4G and future 5G cellular networks.

Future standardization will go in two main directions. First, existing Release 1 standards will be revised, for example, to improve DCC, specify performance requirements, enhance data dissemination concepts, and enable other communication media in addition to ITS-G5. Second, future C-ITS applications will introduce new, more demanding requirements beyond those of the safety warning and awareness applications in Release 1. Some Release 2 activities have already started, such as standards for electro-mobility support. It is foreseen that future standardization activities in ETSI TC ITS will focus on cooperative advanced cruise control (C-ACC), platooning, and protection of vulnerable road users such as pedestrians.

References

[1] Car-2-Car Communication Consortium, http://www.car-2-car.org.
[2] Amsterdam Group, http://www.amsterdamgroup.eu.
[3] CEN and ETSI, Final Joint CEN/ETSI-Progress Report to the European Commission on Mandate M/453; http://www.etsi.org/technologies-clusters/technologies/intelligent-transport, July 2013.
[4] EC, New Connected Car Standards Put Europe into Top Gear, http://europa.eu/rapid/press-release_IP-14-141_en.htm, Feb. 2014.
[5] T. Kosch *et al.*, "Communication Architecture for Cooperative Systems in Europe," *IEEE Commun. Mag.*, vol. 47, no. 5, May 2009, pp. 116–25.
[6] J. B. Kenney, "Dedicated Short-Range Communications (DSRC) Standards in the United States," *Proc. IEEE*, vol. 99, no. 7, July 2011, pp. 1162–82.
[7] E. Ström, "On Medium Access and Physical Layer Standards for Cooperative Intelligent Transport Systems in Europe," *Proc. IEEE*, vol. 99, no. 7, July 2011, pp. 1183–88.
[8] C. F. Mecklenbräuker *et al.*, "Vehicular Channel Characterization and Its Implications for Wireless System Design and Performance," *Proc. IEEE*, vol. 99, no. 7, July 2011, pp. 1189–1212.
[9] A. Festag, P. Papadimitratos, and T. Tielert, "Design and Performance of Secure Geocast for Vehicular Communication," *IEEE Trans. Vehic. Tech.*, vol. 59, no. 5, Mar. 2010, pp. 2456–71.
[10] M. Gramaglia *et al.*, "IPv6 Address Autoconfiguration in GeoNetworking-Enabled VANETs: Characterization and Evaluation of the ETSI Solution," *EURASIP J. Wireless Commun. and Networking*, vol. 2012:19, Jan. 2012, pp. 1–17.
[11] V. Sandonis *et al.*, "Vehicle to Internet Communications Using the ETSI ITS GeoNetworking Protocol," *Trans. Emerging Tel. Tech.*, Oct. 2014.
[12] P. Papadimitratos *et al.*, "Secure Vehicular Communications: Design and Architecture," *IEEE Commun. Mag.*, vol. 46, no. 11, Nov. 2008, pp. 100–09.
[13] J. B. Kenney, G. Bansal, C. E. Rohrs, "LIMERIC: A Linear Message Rate Control Algorithm for Vehicular DSRC Systems," *Proc. ACM VANET*, Las Vegas, NV, Sept. 2011, pp. 21–30.
[14] T. Tielert *et al.*, "Design Methodology and Evaluation of Rate Adaptation based Congestion Control for Vehicle Safety Communications," *Proc. IEEE VNC-Fall*, Amsterdam, Netherlands, Nov. 2011, pp. 116–123.

Biography

Andreas Festag is a research group leader and lecturer at the Technical University of 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. As a researcher, he worked with the Telecommunication Networks Group (TKN) at the Technical University of Berlin, Heinrich-Hertz-Institute (HHI) in Berlin, and NEC Laboratories in Heidelberg, where he last had the position of chief researcher. 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. He has served as Chairman of ETSI Technical Committee ITS WG3 (Networking and Transport) since its creation in 2007.

Implementing Virtual Traffic Lights with Partial Penetration: A Game-Theoretic Approach

Ozan K. Tonguz, Wantanee Viriyasitavat, and Juan M. Roldan

ABSTRACT

Virtual traffic lights (VTL) is a new technology that holds the promise of revolutionizing traffic control in urban areas. The original VTL idea was based on 100 percent penetration of VTL technology. In this article, contrary to conventional wisdom, it is shown that 100 percent penetration is not a necessary condition for deploying VTL technology as it can be implemented with partial or low levels of penetration. Furthermore, based on game-theoretic arguments, it is shown that the adoption of VTL technology can be accelerated by providing incentives to vehicles equipped with VTL technology.

INTRODUCTION

A new technology known as virtual traffic lights (VTL) was recently proposed as a new paradigm for traffic management. VTL is an important efficiency application of vehicular ad hoc networks, especially with the advent of autonomous vehicles [1, 2]. This self-organizing new technology is based on vehicle-to-vehicle (V2V) communications at intersections. VTL can revolutionize traffic management in urban areas as it can substantially reduce the commute time of urban workers, increase productivity, reduce the carbon footprint of vehicles, and lead to a greener environment [3–8]. In a VTL environment, vehicles self-organize to elect a leader, which serves as a "virtual" traffic light (as opposed to a physical traffic light) to decide the right of way at that intersection, thus replacing current physical traffic lights [9–11].

Implementing VTL technology with partial penetration, however, is an outstanding issue that needs to be addressed. V2V communications using dedicated short range communications (DSRC) in the 5.9 GHz band have facilitated the development of VTL [12, 13]. However, given that the introduction of DSRC and VTL technology might occur gradually (with initial penetration rate of only 5–10 percent), systems, methods, and policies must be designed to pave the way for a smooth transition from the current infrastructure-based traffic control system to a full-blown VTL system.

In this article, a new method and system is presented for the coexistence of VTL with the current infrastructure-based traffic control systems under very low penetration rates of DSRC and VTL technology. Note that since the VTL technology requires vehicles to be equipped with DSRC technology, the VTL penetration rate (i.e., the fraction of vehicles that are equipped with VTL technology) is always less than or equal to DSRC radios' penetration rate. In this article, unless specified otherwise, VTL vehicles refer to vehicles that are equipped with both DSRC and VTL, and non-VTL vehicles are vehicles that are either not equipped with DSRC or equipped with DSRC technology but *not* VTL. The presented approach also shows how the transition from current traffic control systems to VTL systems can be made as the penetration rate of DSRC and VTL increases over time to much higher values.

PROBLEM STATEMENT

The original VTL idea assumes 100 percent penetration of DSRC and VTL technology [6]. Although such high penetration rates might be possible in the future, it is clear that a transition from the current physical traffic light system to a VTL system might not happen instantaneously. Such an abrupt change might only happen if the Departments of Transportation (DoTs) of different countries pass new legislation or federal policies making VTL the new standard for traffic control. While this might happen in some countries, it is unrealistic to expect this to happen instantaneously on a global scale. Thus, it is a very critical technical issue to see whether VTL can still be implemented with low penetration rates initially and even afterward with less than 100 percent of vehicles equipped with VTL technology. This is the fundamental technical issue addressed by this article.

As with many technologies, including time-division multiple access/Long Term Evolution (TDMA/LTE) and digital broadcast television, it seems very difficult, if not impossible, to have

Ozan K. Tonguz and Juan M. Roldan are with Carnegie Mellon University.

Wantanee Viriyasitavat is with Mahidol University.

 0163-6804/14/$25.00

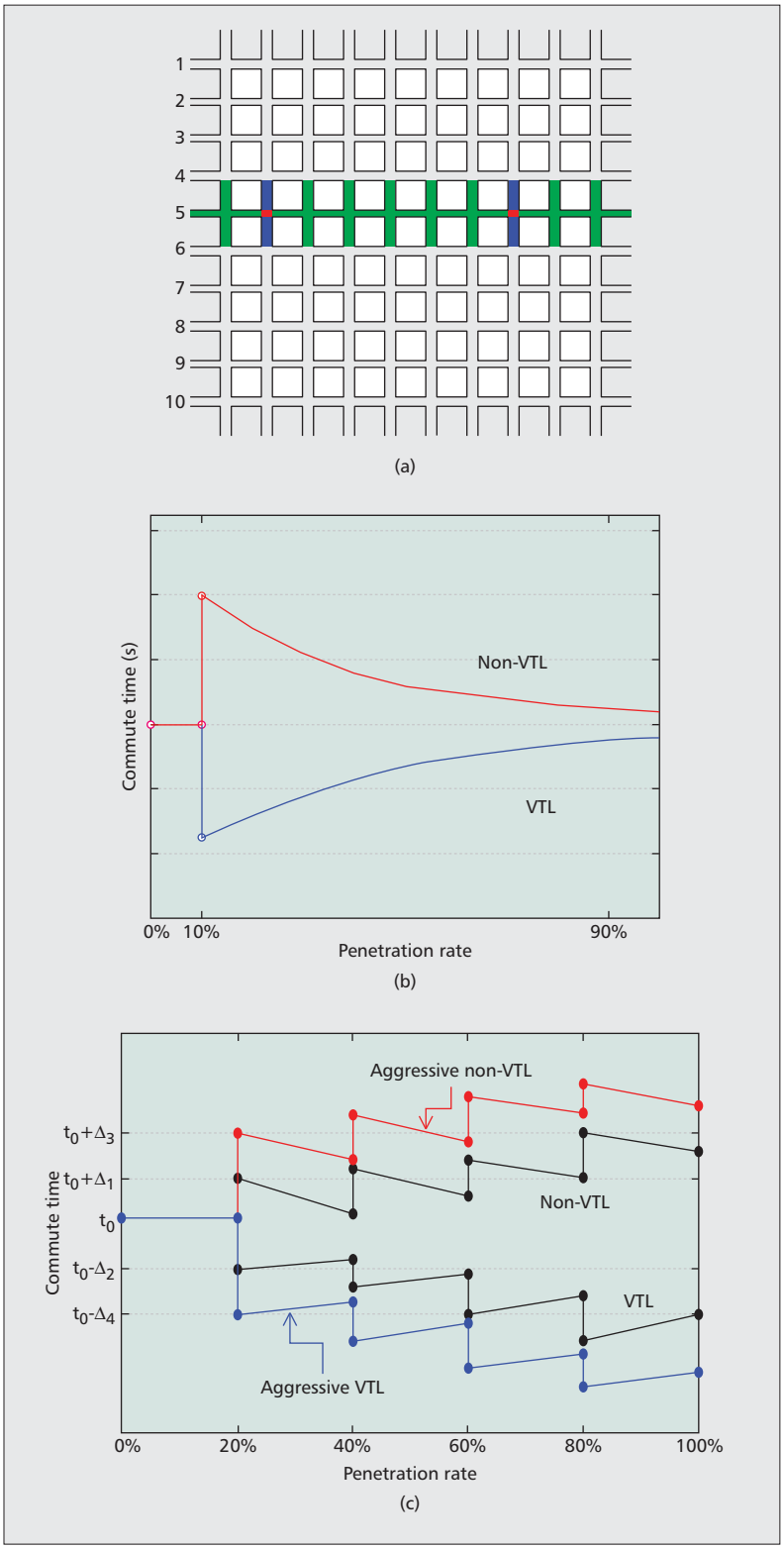


Figure 1. a) The proposed solution for the coexistence model proposed in this article. A route allocated exclusively for VTL vehicles is shown in green color with the exception of two "bridges" (shown in blue color) through which the non-VTL equipped vehicles can cross. In addition to the dedicated route, the VTL vehicles can use any streets in the MG topology; b) a conceptual figure that shows expected result in terms of commute time if the number of routes allocated to VTL vehicles **does not** change as the penetration ratio increases; c) a conceptual figure that shows the expected result in terms of commute time of VTL and non-VTL vehicles when the number of routes allocated to VTL vehicles increases as a function of the increasing penetration rate.

instantaneous adoption of a new technology. In most cases, a transition period to replace the former technology is needed and is crucial for almost all of the new technologies that become commercially viable. DSRC and VTL are no exception to this trend.

In this article, we outline an approach to the implementation of VTL technology in the presence of a very small penetration ratio. The proposed approach is based on game-theoretic concepts. The pioneering work of John von Neumann and Oskar Morgenstern on game theory followed by the important work of John Nash paved the way for using game theory in economics, politics, psychology, computer science, engineering, and biology. Here we use game theory as the underlying basis of our approach. It is shown that using the proposed approach and system during rush hours can expedite the adoption of VTL technology by rewarding VTL vehicles (and their drivers/passengers) while penalizing non-VTL vehicles (and their drivers/passengers).

Solution and the Intuition Behind It

The basic system proposed by this article is simple in nature and depicted in Fig. 1a for the case of low penetration ratio of DSRC and VTL technology (e.g., 5–10 percent of all vehicles are equipped). Observe that the proposed system uses a a dedicated route exclusively reserved for VTL vehicles (shown in green color in Fig. 1a) on which all the infrastructure-based traffic lights are nullified by the VTL-equipped vehicles with the exception of two "bridges" (shown in blue color) through which the non-VTL-equipped vehicles can cross from the upper part of the Manhattan grid (MG) topology to the lower part or vice versa. On such bridges, the functionality of the existing infrastructure-based traffic lights (shown in red) will be maintained. Note that VTL-equipped vehicles are allowed to use any street in the network, while the non-VTL-equipped vehicles can only use the streets that are not reserved for VTL-equipped vehicles and the two bridges as shown in *grkau* and blue, respectively, in Fig. 1a.

It is important to understand what happens if the DoT does *not* change the number of routes allocated to VTL vehicles as the penetration ratio increases. For example, suppose that at 10 and 90 percent penetration the same two routes are allocated to VTL vehicles. Then one would expect to see the trends shown in Fig. 1b for the commute times of VTL and non-VTL vehicles. Note that the graph in Fig. 1b only serves as a conceptual figure and is not drawn to scale; that is, the nonlinear rates of increase or decrease in terms of commute time shown in Figs. 1b and 1c do not reflect actual numbers; rather, they represent the underlying trends. These rates, however, can be designed in a judicious manner by the local DoT authority in a given city.

This is because if two routes are allocated to VTL vehicles at a penetration rate of 10 percent, when the penetration rate increases to 50 percent, for example, the number of vehicles using

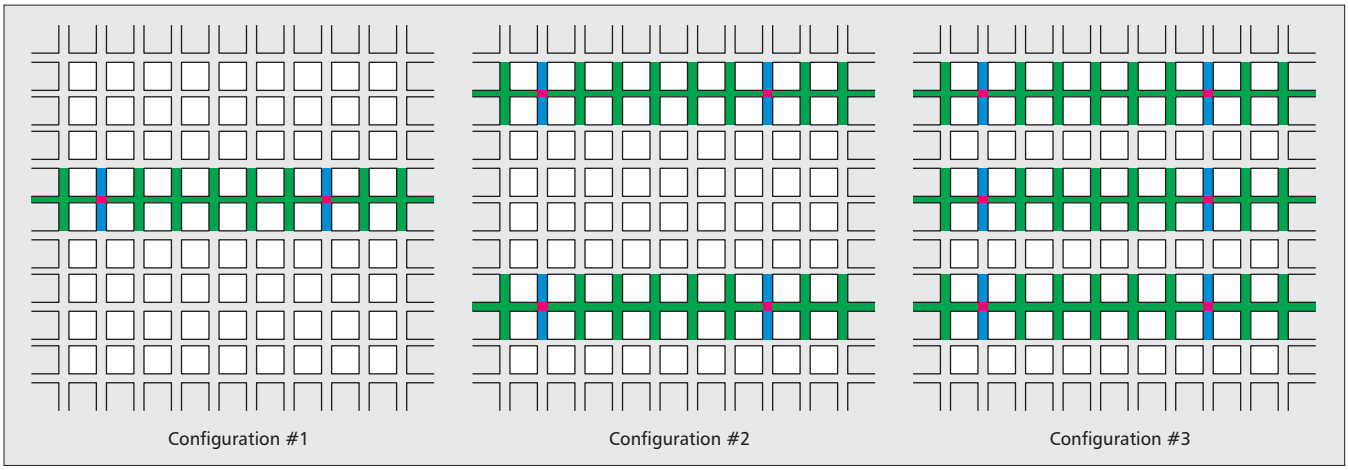


Figure 2. An example of the three configurations of the VTL coexistence model considered in the study. VTL-exclusive streets are highlighted in green color and the streets highlighted in blue color indicate the streets where the non-VTL vehicles may use to cross the VTL-exclusive streets.

those two routes will increase substantially, which, in turn, will lead to more congestion. Note that the 10 percent penetration ratio is assumed to be the first decision point for the DoT to designate a VTL-exclusive route in this article. However, this is clearly a design choice and based on different incentive mechanisms and other factors, different DoTs might decide to do this at different initial penetration ratios. Hence, the commute time of VTL vehicles will gradually start increasing, while the commute time of non-VTL vehicles will gradually start decreasing because far fewer non-VTL vehicles will be using the same road capacity at 50 percent VTL penetration ratio in comparison to 10 percent VTL penetration ratio. This means that the commute time of non-VTL vehicles will decrease after 10 percent penetration if no additional new dedicated routes are allocated to VTL vehicles (this would be very similar to HOV lanes).

In terms of the final results, we expect to obtain the trends shown in Fig. 1c: when penetration rate is 0 percent, all vehicles have a commute time of t_0 s (arbitrary number in Fig. 1c) for the 10 × 10 MG topology we are considering. Assume that in the next two years, 20 percent of vehicles are equipped with DSRC and VTL. The non-VTL vehicles will suffer when the DoT assigns one or two routes exclusively for VTL vehicles; thus, the commute time of non-VTL vehicles increases from t_0 to $t_0 + \Delta_1$. Similarly, the commute time of VTL vehicles will decrease from t_0 to, say, $t_0 - \Delta_2$. If the number of routes allocated to VTL vehicles does not increase, observe that the commute time of non-VTL vehicles will start decreasing from $t_0 + \Delta_1$, while the commute time of VTL vehicles will increase (i.e., get worse) from the initial value of $t_0 - \Delta_2$.

Suppose the penetration ratio reaches 40 percent at the end of the second year, at which point the DoT decides to increase the number of routes allocated to VTL from two routes to three routes. It is important to understand that the number of routes chosen at the end of the second year cannot be arbitrary: it has to be such that for VTL vehicles the new commute time, T, should obey

$$T < t_0 - \Delta_2$$

and for the non-VTL vehicles

$$T > t_0 - \Delta_1.$$

This implies that the policy decision in terms of the number of routes (and which routes are chosen) should be such that the commute times of VTL vehicles get progressively better, while the commute time of non-VTL vehicles get progressively worse.

The rest of the behavior of VTL and non-VTL vehicles shown in Fig. 1c follows the same pattern. As the penetration rate approaches 100 percent, almost all routes are allocated to VTL vehicles, thus rewarding the VTL vehicles by reducing their commute time substantially (around 30–60 percent) while penalizing the non-VTL vehicles by increasing their commute time substantially. Thus, the policy applied by DoT forces non-VTL vehicles to adopt the VTL technology more and more strongly as the penetration rate increases (e.g., through the purchase of the aftermarket VTL devices that will be available on the market).

The red and blue curves in Fig. 1c indicate what happens when the federal government and the DoT decide to apply a more aggressive policy. For example, when the DoT decides to allocate two VTL-exclusive streets (as opposed to one) at 20 percent penetration rate, the average commute time of VTL vehicles will decrease from t_0 to $t_0 - \Delta_4$ which is smaller than $t_0 - \Delta_2$ which is the average commute time of VTL vehicles when there are fewer VTL-exclusive streets. Nevertheless, by allocating additional VTL-exclusive streets, the aggressive policy further penalizes non-VTL vehicles. Observe that when a more aggressive policy is used, the average commute time of non-VTL vehicles substantially increases from t_0 to $t_0 + \Delta_3$ (where $t_0 + \Delta_3 > t_0 + \Delta_1$). In addition, when the penetration rate increases from 20 to 40 percent, the DoT could decide to increase the number of roads allocated to VTL from two to four or two to five, as

The presented approach allows the design of optimum configurations for the number and pattern of VTL-exclusive routes based on penetration ratio, mobility pattern, number of years desired for the adoption of VTL technology, and so on.

opposed to from two to three, to implement such an aggressive policy. Clearly, this will depend on in how many years the federal government and the DoT would like to enforce the migration to VTL technology for traffic management. If the DoT wants to implement this migration in five years as opposed to 10, the policy it needs to apply should be much more aggressive.

Details of the Proposed Approach

Proposed Coexistence Methodology

The methodology used for the coexistence of VTL-equipped vehicles and non-VTL-equipped vehicles is to allocate a major route (or several routes) to only VTL-equipped vehicles during designated times of the day (e.g., during rush hours). Of course, this methodology can only be implemented by the DoT or local (state level) DoTs as a result of a policy decision to provide incentives for the use of VTL technology. Such a policy decision is based on a game-theoretic approach and achieves two important outcomes:

- It reduces the commute time of VTL-equipped vehicles (reward).
- It increases the commute time of non-VTL-equipped vehicles (penalty).

Important Parameters

Intuitively, it is clear that there is a correlation between the penetration ratio of VTL-equipped vehicles and the number of routes exclusively reserved for VTL-equipped vehicles. As the penetration ratio of VTL-equipped vehicles increases, the number of routes exclusively reserved for VTL-equipped vehicles should also increase. This is depicted in Fig. 2.

The other important parameters in the system proposed by this article are the location and number of "special bridges" on the VTL-exclusive routes. Intuitively, it is clear that the number of such crossover points is very critical and could adversely affect the commute time of non-VTL-equipped vehicles if the number of such crossover points is very small. This implies that the number of these crossover points is an important design parameter that could be used to enforce strong penalties on non-VTL-equipped vehicles, thus forcing them to adopt VTL technology in a much faster timeframe.

The Role of Mobility Pattern

The role of the underlying mobility pattern should be taken into account to ensure that the choice of VTL-exclusive routes is made accurately. In the Results section we provide two different case studies as examples of how mobility pattern affects the choice and pattern used for picking VTL-exclusive routes.

Transition as VTL Penetration Ratio Increases

Another important aspect of the presented approach is how to synchronize the choice of VTL-exclusive routes to the increasing penetration ratio. The intuition described previously sheds light on how this issue should be handled. It is very clear that the choice of the number of VTL-exclusive routes for 20 percent penetration ratio is not going to work at 40 percent penetration, as more VTL-exclusive routes will be needed at 40 percent penetration.

Our study shows that the key design issues in this respect are the following:

- The choice of new VTL-exclusive routes should progressively reduce the commute time of VTL vehicles, while it should progressively increase the commute time of non-VTL vehicles.
- The rate of decrease/increase of commute times can be designed to be commensurate with the desired time for the enforced policy to adopt VTL technology (e.g., if a two-year period is envisioned as opposed to a four-year period, a more aggressive policy in the form of number and pattern of VTL-exclusive routes can be applied).

Optimum Configurations

The presented approach allows the design of optimum configurations for the number and pattern of VTL-exclusive routes based on penetration ratio, mobility pattern, number of years desired for the adoption of VTL technology, and so on.

To elaborate on this further, it is clear that optimum configurations should take into account the underlying mobility pattern. For instance, in the optimum configuration, routes that should be allocated exclusively for VTL vehicles should be different when two different mobility patterns, as in cases 1 and 2 in Fig. 3, are used.

Results

The foregoing ideas for implementing VTL technology with partial or low penetration have been assessed via extensive simulations. Below, we describe the conducted simulations and the results obtained.

Simulation Settings

We resort to a simulator of urban mobility (SUMO) traffic simulator to study the trade-off mentioned in the previous section. A 30×30 MG network with two-way one-lane streets is assumed in the simulations, and three configurations with different numbers of VTL-exclusive streets are considered. The three configurations are shown in Fig. 4. All intersections in the network (besides the intersections on the VTL-exclusive streets highlighted in green) are equipped with pre-timed traffic lights with 50-s duration and 50/50 green split. Details of the simulation settings are shown in Table 1. In all of the configurations considered, unless specified otherwise, a unidirectional mobility pattern, as shown in case 1 of Fig. 3, is assumed. This mobility pattern facilitates the assessment of the previously proposed solution. Subsequently, the validity of the proposed solution is evaluated using more realistic mobility patterns as well (e.g., using a radial mobility pattern as depicted in case 2 of Fig. 3).

In the simulations, the leftmost 30×3 area is considered as the source area, and all vehicles

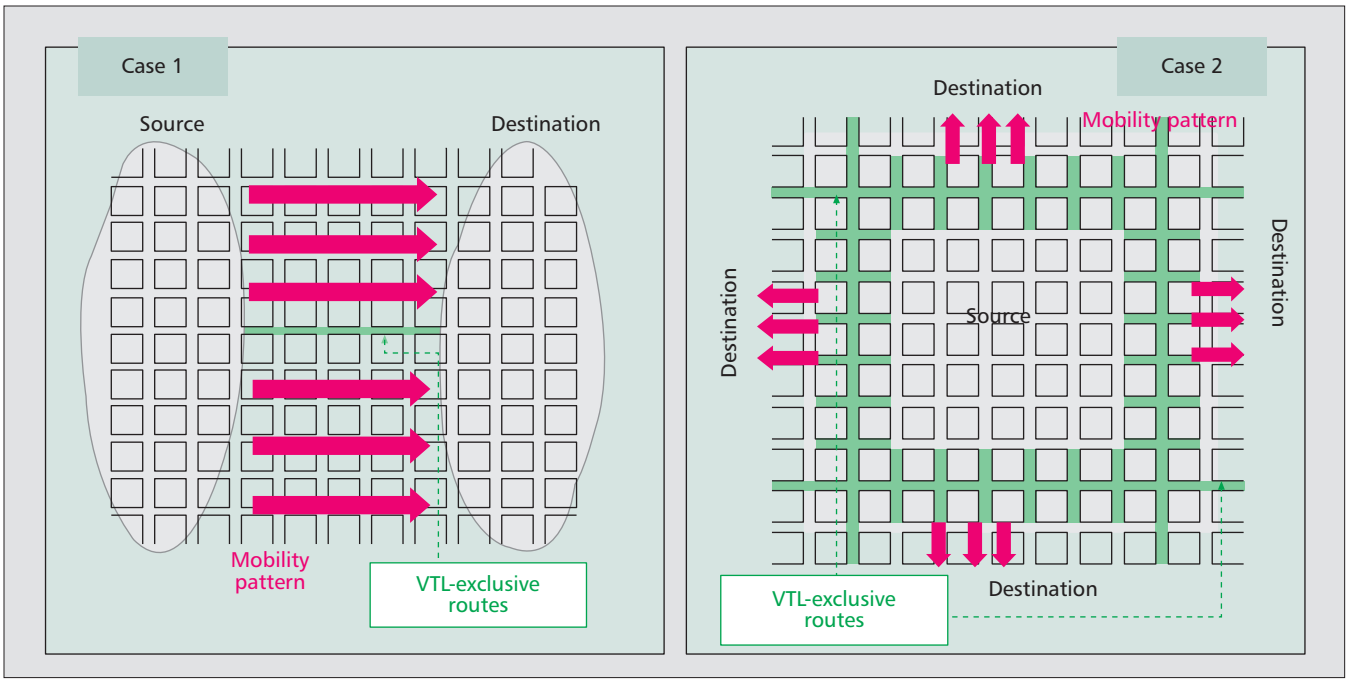


Figure 3. Two different mobility patterns and how they affect the choice of VTL-exclusive routes. Case 1 depicts a unidirectional mobility pattern that is eastbound, whereas case 2 depicts a radial mobility pattern where vehicles move outward from a downtown area.

have a destination location in the rightmost 30 × 1 destination area. Three vehicles are generated each second for a period of 1500 s. In addition, we also assume there are nine main roads connecting the source area to the destination area. The non-VTL vehicles choose to travel on one of the main roads not allocated for VTL vehicles. For instance, in configuration #3, the non-VTL vehicles only use minor roads to travel from the source area to the destination area since all the main roads are VTL-exclusive.

For all the simulations conducted (with different road configurations and penetration ratios), source-destination pairs are generated and fixed. While the source-destination locations of vehicles are fixed, routes taken by these vehicles may be different for a given penetration ratio and road configurations. For instance, in the configuration shown in Fig 4a, the VTL-equipped vehicles tend to choose one of the three main exclusive streets to move toward the destination, while the non-VTL vehicles will choose one of the other six main roads to travel toward the destination. As the number of VTL-exclusive streets increase from three to six as shown in Fig. 4b, the VTL vehicles will tend to use the six exclusive streets, whereas the non-VTL vehicles will choose one of the three remaining main roads (out of a total of nine main roads). It is worth mentioning here that the routes taken by both VTL and non-VTL vehicles are iteratively computed and optimized to minimize the total commute time for all vehicles. This is provided by the SUMO's dual-iterate tool [14]. Note that while the network topology assumed here might not resemble the actual road topology in real cities, it serves as a proof-of-concept example in our study to determine the feasibility of the proposed VTL coexistence model.

Simulation Results

The results presented in this subsection assume two different mobility patterns:

- The eastbound mobility pattern depicted in Case 1 of Fig. 3
- The outward radial mobility pattern depicted in Case 2 of Fig. 3

Figures 4a–c correspond to the unidirectional eastbound mobility pattern depicted as case 1 in Fig. 3. This mobility pattern assumes that the source and destination areas of vehicles are clearly separated. Observe that this mobility pattern could be used to represent the traffic during rush hours. For example, during the morning rush hour, people drive from a residential area (i.e., the source area) to the downtown area (i.e., the destination area). Note that in Figs. 4a–c, the VTL-exclusive roads and crossover points are shown in green and blue colors, respectively. More specifically, observe that crossover points are located at the intersections where the vertical roads of the columns 2, 8, 14, 20 and 26 intersect the 9 main horizontal roads depicted in green color in Figs. 4a–c.

Figure 4d shows the performance of VTL and non-VTL vehicles in terms of average commute time as a function of the penetration ratio of VTL-equipped vehicles. The configuration assumed is shown in Fig. 4a, which shows that at 5 percent penetration three routes are exclusively reserved for the use of VTL vehicles (shown in green) with 15 crossover points (shown in blue) at which infrastructure-based traffic lights are utilized to make sure that non-VTL vehicles can traverse the VTL-exclusive route from top to bottom or vice versa.

Figure 4d shows that using a VTL-exclusive

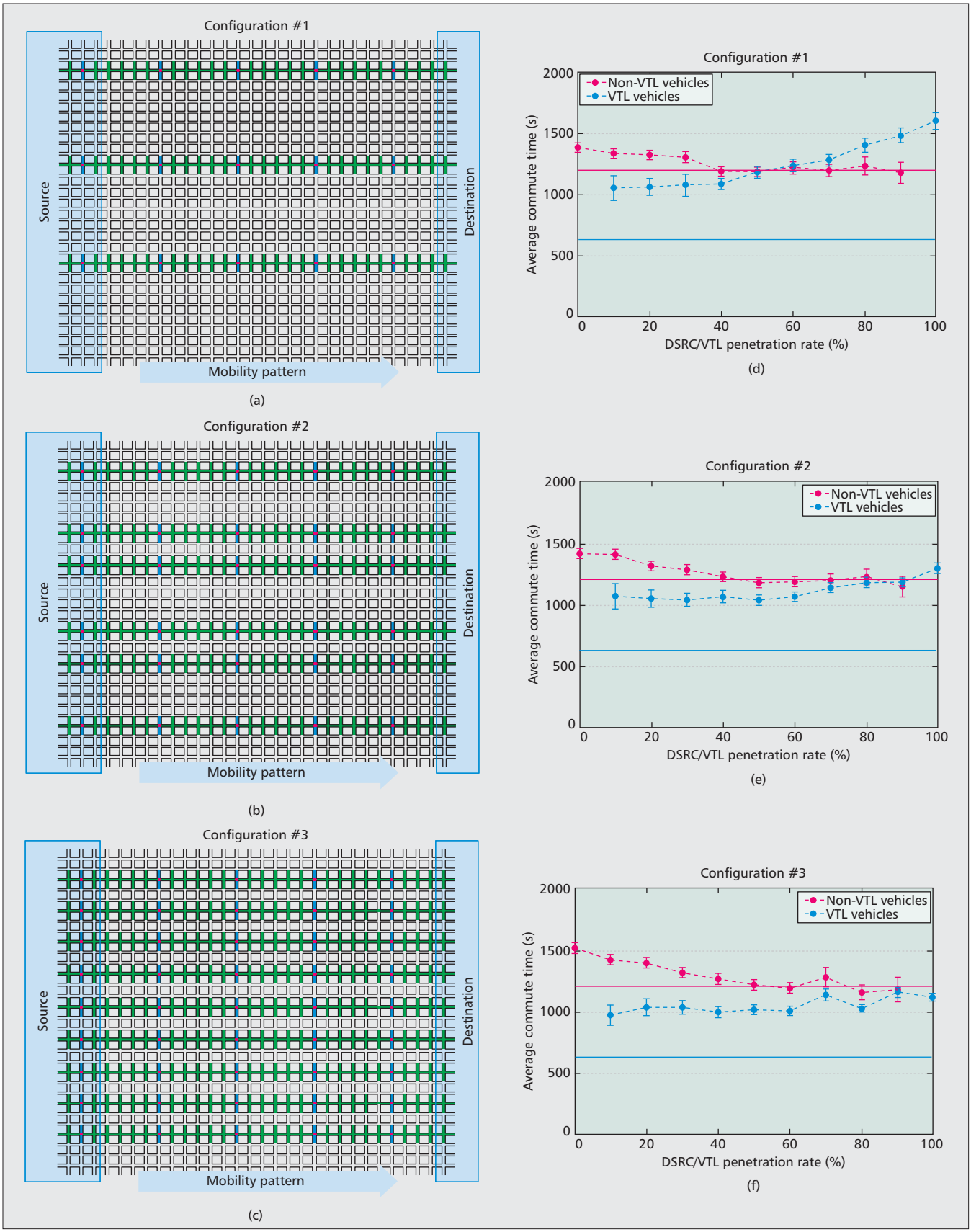


Figure 4. Topology, different configurations, and the corresponding simulation results in terms of the commute time of VTL and non-VTL vehicles shown as a function of penetration ratio: a) three routes (shown in green) are exclusively reserved for the use of VTL vehicles with 15 crossover points (shown in blue); b) six routes are exclusively reserved for the use of VTL vehicles with 30 crossover points; c) nine routes are exclusively reserved for the use of VTL vehicles with 45 crossover points; d) shows the results corresponding to the configuration in a); e) shows the results corresponding to the configuration in b); f) shows the results corresponding to the configuration in c), respectively. The results are plotted with 95 percent confidence intervals.

Parameters	Values
Topology used	30 × 30 MG
Length of a block	125 m
Number of streets	60
Number of lanes in each direction	1
Green split	20 s
Yellow split	5 s
Cycle duration	50 s
Traffic generation rate	3 cars/s

Table 1. Parameter values used in the simulation.

route can reduce the commute time of VTL vehicles (shown in blue) with respect to the commute time of non-VTL vehicles (shown in red). The horizontal blue line represents the commute time (629 s) with 100 percent VTL penetration, while the red horizontal line represents the commute time (1198 s) with 0 percent VTL penetration. Observe that reserving a VTL-exclusive route at 10 percent is very effective in giving an edge/advantage to VTL vehicles initially. More specifically, Fig. 4d shows that with the proposed approach, at 10 percent penetration ratio, while the average commute time of VTL vehicles is reduced from about 1200 to 1053 s (which is a 12 percent decrease), the average commute time of non-VTL vehicles increases to 1334 s (which is an 11 percent increase). Thus, the differential between the commute times of VTL and non-VTL vehicles is 281 s, which corresponds to a commute time for non-VTL vehicles that is 27 percent higher than that of the VTL vehicles. Clearly, this is a significant difference. The observed benefit, however, reduces as the penetration ratio of VTL vehicles increases to 100 percent. In fact, Fig. 4d clearly shows that the commute time of VTL and non-VTL vehicles converges to approximately the same value as the penetration ratio approaches 50 percent (the figure shows that at 50 percent the commute time of VTL vehicles becomes slightly worse than non-VTL vehicles). The disappearance of the polarization between VTL and non-VTL vehicles in terms of commute time confirms intuition in the sense that as the penetration ratio of VTL vehicles increases, using the same VTL-exclusive routes is no longer effective as it leads to a major bottleneck and congestion for VTL vehicles. This phenomenon also points to the significance of using a *dynamic pattern* for VTL-exclusive routes, which should increase the number of routes used for VTL vehicles as the penetration ratio increases.

Another important observation that can be made is the drastic reduction of commute time with 100 percent VTL penetration (blue horizontal line — 629 s) with respect to the commute time with 0 percent VTL penetration (red horizontal line — 1198 s). This corresponds to a 48 percent reduction in commute time, which shows the substantial improvement that could be achieved with VTL technology. This is in line with the previous results reported on the benefit of VTL [6].

One can also quantify the benefit of using a VTL-exclusive route by using the speed of VTL and non-VTL vehicles. While these results are omitted due to lack of space, we have observed that the same polarization takes place, and while the average speed of VTL vehicles increases, the average speed of non-VTL vehicles decreases. If one keeps the same static arrangement of a single VTL-exclusive route, as expected, the speeds of VTL and non-VTL vehicles converge as penetration ratio tends to 100 percent.

Figure 4e shows the results obtained from the same MG scenario with the same eastbound mobility pattern when one uses six VTL-exclusive routes as opposed to three. Observe that, for this case, the commute time of VTL vehicles can be reduced to lower values (1030 s at 50 percent VTL penetration) with respect to the commute time in the previous configuration (around 1180 s at the same penetration level) in Fig. 4d. The figure also shows that the polarization between VTL and non-VTL vehicles is more persistent as they converge at higher values of penetration ratio (around 80 percent) with respect to the previous configuration.

Figure 4f shows the simulation results when one uses nine VTL-exclusive routes for the same topology and mobility pattern (as shown in Fig. 4c) at 5 percent penetration. One can observe that the results for this configuration, as expected, are even better than for the previous two cases. In fact, with this configuration no convergence in the commute times of VTL and non-VTL vehicles occur even at 100 percent penetration.

Figure 5 shows the simulation results of similar topology of a 10 × 10 MG with a *different mobility pattern*. In particular, the results in Fig. 5 assume the source area to be in the middle, and the mobility pattern to be radial and outward, as shown in case 2 of Fig. 3. More specifically, in this case it is assumed that the vehicles originate from a point in the 9 × 9 inner square grid and move toward their destinations located on the boundary of the 10 × 10 MG network. Observe that because of the underlying radial mobility pattern, locations of the VTL-exclusive roads are different from those assumed previously. As a result, the configurations of VTL-exclusive streets are redesigned in such a way that it is symmetric along both the x- and y-axes. The VTL-exclusive routes are shown in the figure in green, and a dynamic strategy is utilized whereby the patterns used in the 20–40 percent range is configuration A, in the 40–60 percent range configuration B, and in the 60–80 percent range configuration C. Observe that as the penetration ratio increases, a larger number of VTL-exclusive routes are utilized, which is the right strategy.

The commute times of VTL vehicles are shown in blue, while the commute times of non-VTL vehicles are shown in red. Observe that the results obtained confirm the intuition described

In this article, contrary to conventional wisdom, it is shown that 100 percent penetration is not a necessary condition for deploying VTL technology. In fact, the results presented clearly show that VTL can be deployed even at very low levels of penetration, which seems counter-intuitive.

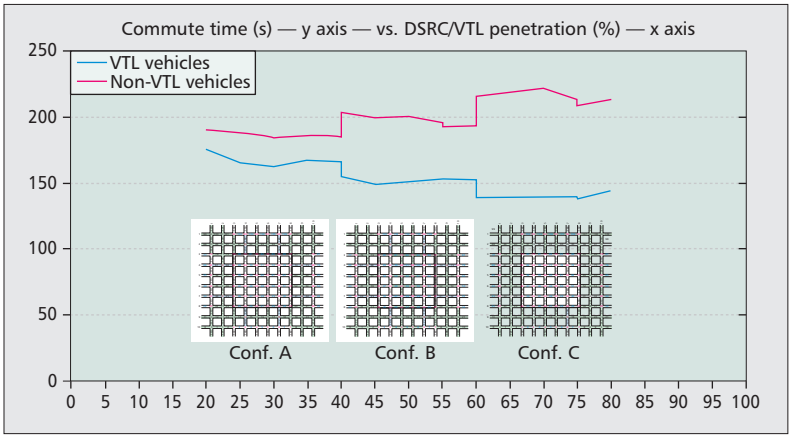


Figure 5. Simulation results in terms of commute time of VTL and non-VTL vehicles shown as a function of penetration ratio for three different scenarios/configurations. In this scenario, a radial mobility pattern is assumed in the simulations whereby vehicles originate from a point in the 9×9 inner square grid (case 2, Fig. 3) and move toward the network boundary. Routes (shown in green) are exclusively reserved for the use of VTL vehicles with two crossover points (shown in blue).

earlier. More specifically, these results show that by using a dynamic strategy that uses a number and pattern for VTL-exclusive routes commensurate with the penetration ratio, one can provide strong incentives to drivers to adopt VTL technology. Observe that by using such a dynamic strategy, one can reinforce the underlying polarization (i.e., the reward and penalty), thus providing a compelling case for the adoption of VTL technology.

Discussion

To put the results presented in this article into perspective, it is important to understand that the original VTL idea was based on the assumption that all vehicles in an urban area will be equipped with VTL technology. In this article, contrary to conventional wisdom, it is shown that 100 percent penetration is not a necessary condition for deploying VTL. In fact, the results presented clearly show that VTL can be deployed even at very low levels of penetration, which seems counter-intuitive.

More specifically, we have presented a road map that shows how, starting with very low levels of penetration, the use of VTL can be supported by reserving certain routes during certain times (e.g., rush hours) to the exclusive use of VTL-equipped vehicles. The consequence of this approach is to reward VTL-equipped vehicles by reducing their commute time while penalizing non-VTL equipped vehicles by increasing their commute time. This game-theoretic approach holds the promise of not only supporting the deployment of VTL technology with partial penetration but even accelerating it, similar to well-known reinforcement algorithms in artificial intelligence.

While a comprehensive analysis on commute times in real settings is necessary, the results presented in this article provide guidelines on how to choose VTL-exclusive routes based on the underlying mobility pattern, penetration ratio, and targeted adoption time in a given urban area. The approach presented clearly underscores the significance of federal and local governments in implementing such policies for the partial deployment of VTL and accelerating its adoption. One interesting issue is how to achieve the first 5–10 percent penetration of VTL technology at the initial deployment stage (e.g., Fig. 1c). Here again, incentives or subsidies provided by the federal government will be crucial (e.g., such incentives and subsidies are already available in the United States for solar panels in households based on the energy policy of the federal government).

It is also important to note that the assumption on DoT involvement can be somewhat relaxed as it has been shown recently that VTL technology as well as other VANET applications can also be implemented on smartphones as an app as opposed to using standalone DSRC/VTL units [15, 16]. While this might circumvent the need for using new vehicles equipped with standalone DSRC/VTL units (since after-market devices implementing VTL as a smartphone app could also do the same function), given that smartphone usage might not be 100 percent might still necessitate federal or local government decisions in determining which routes will be reserved for VTL use (e.g., during rush hours).

Open Challenges

VTL was originally proposed in 2010 as a self-organizing traffic control paradigm that could mitigate traffic congestion in major cities. The original VTL idea was based on the assumption that to implement this new technology one would need 100 percent penetration of VTL technology; that is, all vehicles would be equipped with VTL technology. This was considered to be a strong assumption and a major limitation given that DSRC technology was not mandated by the U.S. DoT in 2010.

With the announcement of the U.S. DoT and National Highway Traffic Safety Administration (NHTSA) to mandate DSRC technology in February 2014, the landscape has changed, and now it is clear that all vehicles will be equipped with DSRC radios. This has made the implementation of VTL much easier since ubiquitous use of DSRC radios is a necessary condition for implementing VTL. Nevertheless, the current industry forecasts for the United States and Europe predict the market penetration to reach 40–50 percent by 2025 (i.e., in another decade or so). Hence, it is still not clear whether, with such gradual penetration ratios, it would be practical to implement VTL in the near future.

Implementing VTL under partial (or even small) levels of penetration is therefore a challenge, and this article proposes a road map for reaching this goal. There are several additional challenges that need to be addressed for bringing this revolutionary technology (the VTL invention was recently issued a U.S. patent by the U.S. Patent Office) to the marketplace. Below, we summarize some of these important challenges.

How to Deal with Pedestrians — This is an

important issue as pedestrians should have some rights in crossing the proposed VTL-exclusive routes. Although pedestrian crossings (also known as "zebra crossings") along a road block can use the principle of operation that is currently used where the pedestrians have priority over vehicles by law, managing pedestrians at intersections is not straightforward, and will need new and creative engineering solutions. While some camera-based solutions using computer vision might be possible, it is not clear whether such solutions will be cost effective, easy to implement, and robust to different light, brightness, and weather conditions. Car-centric solutions at intersections whereby the right of way for pedestrians is signalized by vehicles might also be a promising solution. Yet another option might be to keep some infrastructure at intersections designed for signaling pedestrians when to cross an intersection.

How to Incorporate Cyclists on the Proposed VTL-Exclusive Routes — One significant difference between bicycles and pedestrians is the fact that bicycles are considered to be vehicles, whereas pedestrians are not. This suggests that cyclists should have the same VTL equipment as other vehicles and obey the same rules for implementing VTL on the proposed VTL-exclusive routes (VTL leader election, etc.). This has its own set of challenges that need to be carefully studied. For example, equipping bicycles with VTL devices could be nontrivial if one uses standalone VTL units, whereas it will be less of a problem if VTL technology is implemented as an app on smart phones. The speed difference between bicycles and other vehicles could also have important ramifications that should be taken into account.

Fail-Safe Design of Virtual Traffic Lights — This is another open challenge for the proposed approach to implementing VTL with partial penetration. There are several potential issues that could lead to undesired outcomes in the operation of VTL in the proposed approach. Among such issues, the obvious ones are GPS inaccuracies, ambiguities in VTL leader election, propagation problems, hardware failures, and so on. To design a fail-safe VTL system is therefore crucial to make sure that under no circumstances do undesired outcomes (e.g., accidents) occur. Using a finite state machine formalism with the legitimate states and a fail-safe state with the necessary transitions appears to be the right approach for meeting this challenge.

RF Propagation Problems — RF propagation problems at certain intersections could adversely affect the proper operation of VTL in the approach described in this article. How to solve this problem in a cost-effective, simple, and robust manner is another important challenge that needs to be addressed. The proposed approach to implementing VTL with partial penetration assumes that V2V communications can be supported at all intersections. In every urban area, however, there are intersections where obstructions might prevent having V2V communications in a timely manner, which is crucial for proper operation of VTL. This is an important open challenge awaiting practical solutions.

Security — Security is another open challenge for the implementation of the proposed approach in this article. Making sure that certain security attacks cannot cripple the operation of VTL with partial penetration will need innovative and sustainable solutions that could work with the DSRC standard, its bandwidth, its number of channels, data rates, and so on.

Making sure that certain security attacks cannot cripple the operation of VTL with partial penetration will need innovative and sustainable solutions that could work with the DSRC standard, its bandwidth, its number of channels, data rates, and so on.

Related Work

It is interesting to note that, in principle, the approach presented in this article is synergistic to the high occupancy vehicles (HOV) lane concept whereby some lanes on a highway, for instance, are reserved for the exclusive use of vehicles carrying passengers in addition to the drivers of those vehicles [17, 18]. The rationale behind the HOV lane concept is to motivate more and more drivers to do car pooling, thus reducing the number of vehicles during rush hours on a highway, which, in turn, may reduce congestion. However, there are key differences between the HOV approach and the approach described in this article in terms of the choice of VTL-exclusive routes and the number of crossover points on these VTL-exclusive routes, which depend on, among other factors, mobility pattern, penetration ratio, the desired time for the policy to take place, and so on. By allocating VTL-equipped vehicles dedicated routes during rush hours, the federal (or local) government can provide strong incentives for the use of VTL technology that can mitigate congestion drastically.

After the publication of the original VTL idea, the interest in VTL technology has increased considerably [5, 6, 9, 10, 12]. It can be seen from the reported results that the growing interest in VTL spans both academic institutions as well as industry (e.g., the article in [5] was recently published by Audi).

While a recent paper by Fathollahnejad *et al.* from Chalmers University investigates the leader election algorithms for VTL [10], an interesting paper by Neudecker *et al.* from Karlsruhe Institute of Technology looks into how to design VTL in a fail-safe manner [12]. To this end, the paper proposes a formal verification and model checking approach, and shows that VTL can be designed in a completely fail-safe manner at the expense of 2 percent degradation in throughput at a given intersection. The work by Sommer *et al.* from Innsbruck University investigates another crucial aspect of VTL: the impact of a wireless propagation environment and network load due to traffic densities on the performance of the VTL system [11]. The recent work by Neudecker explores how the RF propagation issues affect the feasibility of VTL at challenging intersections that might have high rises or tall buildings at the intersection corners [13].

While all of these recent studies explore different aspects of VTL technology, none of them focuses on investigating whether VTL is feasible without 100 percent penetration and, if so, how

While all of these recent studies explore different aspects of VTL technology, none of them focuses on investigating whether VTL is feasible without 100 percent penetration and, if so, how to deploy VTL with partial penetration. The work presented in this article aims to address this important problem.

to deploy VTL with partial penetration. The work presented in this article aims to address this important problem.

Conclusion

A game-theoretic approach is presented for implementing virtual traffic lights technology with partial or very small penetration ratios. The presented approach facilitates the coexistence of VTL technology with the currently used infrastructure-based traffic control systems. To this end, it is shown that certain routes can be reserved for VTL-equipped vehicles at designated times of the day (e.g., during rush hours). This, in turn, reduces the commute time of VTL vehicles while it increases the commute time of non-VTL vehicles, thus providing strong incentives for accelerating the adoption of VTL technology. Our results also show how the number of VTL-exclusive routes and the crossover points on them can be designed taking into account the penetration ratio, the mobility pattern, the time duration desired for the adoption of VTL technology, as well as other factors.

Acknowledgment

This research was supported in part by the T-SET University Transportation Center sponsored by the U.S. Department of Transportation under Grant No. DTRT12-G-UTC11.

References

[1] H. Hartenstein and K. Laberteaux, "A Tutorial Survey on Vehicular Ad Hoc Networks," *IEEE Commun. Mag.*, vol. 46, no. 6, 2008, pp. 164–71.
[2] M. Gerla, "Quod Vides? or Vehicular Cloud Computing," *Proc. Dagstuhl Wksp., Inter-Vehicle Commun.: Quo Vadis*, Sept. 2013.
[3] W. Ferguson, "Virtual Traffic Lights Help Solve Commuting Hell," *New Scientist Mag.*, Nov. 2012.
[4] E. Jaffe, "How Virtual Traffic Lights Could Cut Down on Congestion," *The Atlantic*, Feb. 2013.
[5] O. Strohbach, "Green Wave," *AUDI's Encounter Tech. Mag.*, issue 7, Sept. 2013, p. 45.
[6] M. Ferreira *et al.*, "Self-Organized Traffic Control," *Proc. ACM Int'l. Wksp. VANET*, 2010, pp. 85–90.
[7] O. K. Tonguz, "Biologically Inspired Solutions to Fundamental Transportation Problems," *IEEE Commun. Mag.*, vol. 49, no. 11, Nov., 2011, pp. 106–15.
[8] W. Viriyasitavat and O. K. Tonguz, "Priority Management of Emergency Vehicles at Intersections Using Self-Organized Traffic Control," *Proc. IEEE VTC-Fall*, 2012, pp. 1–4.
[9] S. Joerer *et al.*, "A Vehicular Networking Perspective on Estimating Vehicle Collision Probability at Intersections," *IEEE Trans. Vehic. Tech.*, vol. PP, no. 99, 2013, p. 1.
[10] N. Fathollahnejad *et al.*, "On rEliability Analysis of Leader Election Protocols for Virtual Traffic Lights," *Proc. IEEE/IFIP Conf. Dependable Sys. and Networks Wksp.*, June 2013, pp. 1–12.
[11] C. Sommer, F. Hagenauer, and F. Dressler, "A Nework Perspective on Self-Organizing Intersection Management," *Proc. IEEE World Forum on Internet of Things*, Seoul, Korea, Mar. 2014, pp. 230–34.
[12] T. Neudecker, N. An, and H. Hartenstein, "Verification and Evaluation of Fail-Safe Virtual Traffic Lights Applications," *Proc. IEEE Vehic. Networking Conf.*, Boston, MA, Dec. 2013.
[13] T. Neudecker, *Feasibility of Virtual Traffic Lights under Realistic Communication Conditions*, MSc. Thesis, Karlsruhe Inst. Tech., May 2012.
[14] M. Behrisch *et al.*, "Sumo — Simulation of Urban Mobility: An Overview," *Proc. SIMUL 2011*, Barcelona, Spain, October 2011, pp. 63–68.
[15] M. Nakamurakare, W. Viriyasitavat, and O. Tonguz, "A Prototype of Virtual Traffic Lights on Android-Based Smartphones," *Proc. IEEE SECON*, New Orleans, LA, June 2013, pp. 236–38.
[16] Y. Park *et al.*, "A Feasibility Study and Development Framework Design for Realizing Smartphone-Based Vehicular Networking Systems," *IEEE Trans. Mobile Computing*, Mar. 2014.
[17] R. G. Dowling *et al.*, "Predicting High Occupancy Vehicle Lane Demand," Federal Highway Administration, U.S. DoT, rep. no. FHWA-SA-96-073, Aug. 1996.
[18] J. Kwon and P. Varaiya, "Effectiveness of California's High Occupancy Vehicle (HOV) System," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 1, 2008, pp. 98–115

Biographies

Ozan K. Tonguz is a tenured full professor in the Electrical and Computer Engineering Department of Carnegie Mellon University (CMU). He currently leads substantial research efforts at CMU in the broad areas of telecommunications and networking. He has published about 300 papers in IEEE journals and conference proceedings in the areas of wireless networking, optical communications, and computer networks. He is the author (with G. Ferrari) of the book *Ad Hoc Wireless Networks: A Communication-Theoretic Perspective* (Wiley, 2006). In December 2010, he founded the CMU startup known as Virtual Traffic Lights, LLC, which specializes in providing solutions to acute transportation problems using vehicle-to-vehicle and vehicle-to-infrastructure communications paradigms. His current research interests include vehicular networks, wireless networks, sensor networks, self-organizing networks, smart grid, bioinformatics, and security. He currently serves or has served as a consultant or expert for several companies, major law firms, and government agencies in the United States, Europe, and Asia.

Wantanee Viriyasitavat is a lecturer in the Faculty of Information and Communication Technology at Mahidol University, Bangkok, Thailand. During 2012–2013, she was a research scientist in the Department of Electrical and Computer Engineering at CMU. She received her B.S./M.S. and Ph.D. degrees in electrical and computer engineering from CMU in 2006 and 2012, respectively. Between 2007 and 2012, she was a research assistant at CMU, where she was a member of the General Motors Collaborative Research Laboratory working on the design of a routing framework for safety and non-safety applications of vehicular ad hoc wireless networks. Her research interests include traffic mobility modeling, network connectivity analysis, and protocol design for wireless ad hoc networks.

Juan M. Roldan is a Ph.D. student in the Engineering and Public Policy Department of CMU. He holds a B.A. in electrical engineering and an M.A. in economics, both from the University of Los Andes, Bogota, Colombia. His current research interests include policies and regulation for the telecommunications sector and policies for the deployment of Virtual Traffic Lights, a new technology pursued by the CMU startup Virtual Traffic Lights, LLC. He has served as a telecommunications consultant for several private and public entities in the United States and Latin America.

Intra-Car Multihop Wireless Sensor Networking: A Case Study

Morteza Hashemi, Wei Si, Moshe Laifenfeld, David Starobinski, and Ari Trachtenberg

ABSTRACT

Modern vehicles incorporate dozens of sensors to provide vital sensor data to electronic control units, typically through physical wires, which increase the weight, maintenance, and cost of cars. Wireless sensor networks have been contemplated for replacing the current physical wires with wireless links, although existing networks are all single-hop, presumably because cars are small enough to be covered by low-power communication, and multihop networking requires organizational overhead. In contradiction with previous works, we experimentally investigate the use of multihop wireless communication to support intra-car sensor networking. Extensive tests, run under various vehicular environments, indicate the potential for significant reliability, robustness, and energy usage improvements over existing single-hop approaches. Our implementation is based on the Collection Tree Protocol, a state-of-the-art multihop data collection protocol.

INTRODUCTION

Wireless sensor networks (WSNs) boast numerous applications ranging from home appliance control to environmental monitoring and smart healthcare. More recently, they have also demonstrated benefits for intelligent transport systems: monitoring aircraft systems and parameters [1], monitoring wheel bearings on trains [2], and connecting sensors, switches, and actuators inside cars [3–5]. In these applications, the ability to reliably aggregate data in one or several processing centers is critical to the monitoring capabilities of the sensors, which are typically constrained in both energy and computational power. For transport systems, this aggregation is further complicated by the dynamic channel properties that vehicles may experience as they travel through areas with different radio interference patterns or road quality.

To date, several *single-hop* communication models based on Zigbee, RFID, and ultra-wideband technologies, have been examined for intra-car wireless networking [3–5]; in these networks, all sensor nodes communicate in a point-to-point fashion (within a star topology) with the central node. For large-scale deployments, especially over large physical distances, it may be advantageous to arrange sensors in a *multihop* network, with some sensors relaying information from other sensors onto the central node (or *collection root*). Within vehicles, however, it is not clear whether multihop networking is worthwhile, as the small distances involved typically allow sensors to reach the central node using low or medium transmission power. Indeed, multihop networking adds communication overhead to the system, requiring exchange of metadata (e.g., topology information) and utilizing available bandwidth for packet relaying (with some incident interference to other sensors).

In this work, we show that, notwithstanding the caveats mentioned above, multihop networking may provide clear benefits within cars. Specifically, despite its greater overhead, it can enhance system reliability, provide robust performance, and reduce communication energy. We utilize the Collection Tree Protocol (CTP) [6] as a multihop network layer protocol, as it is widely deployed, well researched, and the basis of the IPv6 Routing Protocol for Low-Power and Lossy Networks (RPL) standard [7]. Our results show, for example, that the packet delivery rate of a node using a single-hop topology protocol can be below 80 percent in practical scenarios, whereas CTP improves reliability performance beyond 95 percent across *all* nodes while simultaneously reducing radio energy consumption.

We perform several experiments to explore the effects of environmental conditions on the performance of intra-car WSN; for instance, our results show that engine noise does not have much effect on performance, even though it adds 2–4 dB in noise power [3]. On the other hand, CTP achieves clearly better performance when a vehicle is parked in a covered area (i.e., garage) than in an open area (i.e., outdoor parking). To the best of our knowledge, this is the first demonstration of the significant benefits possible from multihop networking within vehicles. A preliminary version of our results was presented in [8].

The rest of this article is organized as follows. In the following section we provide background

Morteza Hashemi, Wei Si, David Starobinski, and Ari Trachtenberg are with Boston University.

Moshe Laifenfeld is with General Motors.

 0163-6804/14/$25.00

The Collection Tree Protocol is a variant of a distance-vector routing protocol that includes optimization tailored for wireless sensor networks. CTP is designed to route data from every node on a network to one or more self-declared root nodes, based on minimum cost trees.

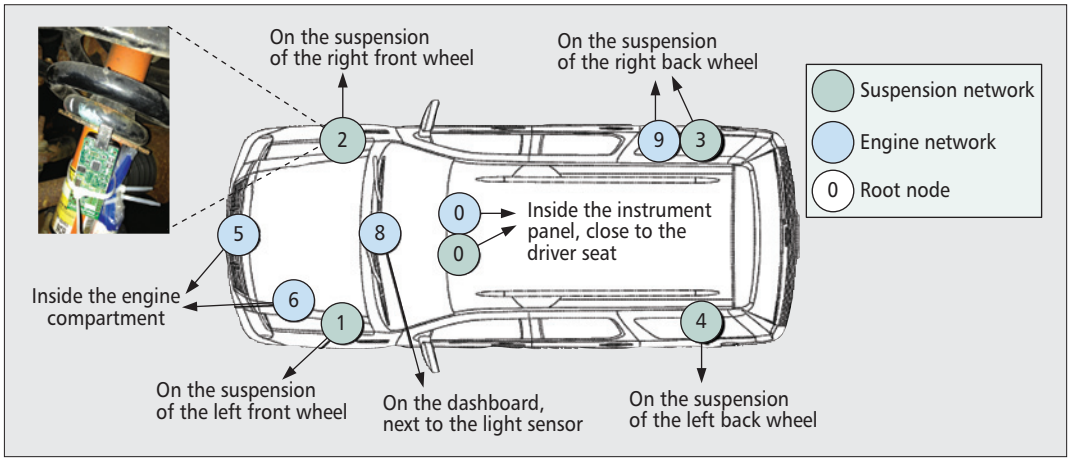


Figure 1. Placement of nodes inside the car for the suspension network and engine network.

on existing intra-car WSN models and on CTP. After that our experimental setup is described. We follow that with a comparison of multihop CTP performance to a star topology protocol under static conditions. Then we extend the experiments into dynamic scenarios. We conclude with overall thoughts.

Background

Intra-Car Wireless Sensor Networking

Several intra-car WSN experiments have been conducted for the 915 MHz and 2.4 GHz industrial, scientific, and medical (ISM) bands, since many off-the-shelf products (ZigBee, Bluetooth, RFID) operate at these frequencies. The authors in [3] characterize the physical layer of a Zigbee-based sensor network inside a car; in particular, they measure the received signal strength indicator (RSSI) and link quality indicator (LQI). The results in [4, 9, 10] provide comprehensive statistics for the intra-car channel, including its power delay profile (PDP), coherence bandwidth, and coherence time. Other fading statistics, such as level-crossing rate (LCR) and average fading duration (AFD), have also been measured in these works. In [5], ultra-wideband (UWB) technology is considered for short-range communication within a car due to its low power requirements and high data rate.

In contrast to previous works, we investigate the performance of an intra-car *multihop* WSN on a commercial TelosB platform. Multihop data collection can potentially compensate for large channel losses. For example, the results in [10] show that the average channel loss of a link between a transmitter (located in the engine compartment) and a receiver (placed inside the trunk) is about 85 dB. Considering the receiver sensitivity of the sensor's radio chip (e.g., the CC2420 has a sensitivity of –94 dBm [typical] and –90 dBm [minimum]), one can conclude that sensor nodes should always transmit at high power to overcome such channel loss. On the other hand, the situation is different for a multihop network, wherein sensor nodes can opportunistically choose the next hop according to ongoing channel conditions and decrease the transmission power while communicating reliably, possibly at the expense of increased latency and decreased network throughput. Our work thus aims to concretely evaluate the cost-benefit regions of multihop WSN inside cars.

Collection Tree Protocol

The Collection Tree Protocol is a variant of a distance-vector routing protocol that includes optimization tailored for wireless sensor networks. CTP is designed to route data from every node on a network to one or more self-declared root nodes, based on minimum cost trees. In [6] a variant of CTP is introduced, in which the expected number of transmissions, denoted by ETX, is used as the cost metric. In this approach, each node that has a message to transmit attempts to build a shortest path tree with a minimum number of transmissions toward the root. The calculation of ETX for each node is done as follows: Consider two nodes, A and B, such that node A is the parent of node B; then we have

$$ETX_B = ETX_A + \text{ETX of link } B \rightarrow A;$$
$$ETX_{Root} = 0,$$

where the ETX of a link is estimated by a link estimator in a distributed fashion [6]. Given a choice of valid routes, CTP simply chooses the one with the lowest ETX toward the root.

Experimental Setup

Experimental Methodology

In the experiments, we evaluate two distinct networks operating on different frequencies inside a car: a *suspension network* and an *engine network*. Each network consists of four sensor nodes periodically sending data to the collection root, which itself is connected to a laptop through a USB port for the purposes of logging messages and statistics for post-experiment analyses. Figure 1 graphically illustrates the placement of nodes in the car. Note that previous studies (e.g., [10]) show that wireless channels to other locations in the vehicle show more or less similar characteristics, that is, slow fading with coherence time of several seconds.

In addition to the sensor nodes and collection root, we utilize an *activator node* (not shown in Fig. 1) to send an initial broadcast signal that activates each sensor node and establishes basic time synchronization. The use of an activation broadcast enables us to dictate the packet generation rate, transmission power, and radio channel for the nodes. The activator node is chosen to be different from the root node because all nodes should receive the activation signal, while the signal of the root may not be received by all nodes under low-power transmission. Therefore, we use the activator node to send a high-power broadcast signal received by all sensors.

Our sensor nodes are TelosB (TPR2420CA) [11] with a data rate of 250 kb/s, RF power –24 dBm to 0 dBm, and receiver sensitivity of –94 dBm (typical) through –90 dBm (minimum). The sensor nodes' firmware is based on TinyOS 2.x, which is an open source operating system designed for sensor networks. The radio chip of sensor nodes (CC2420) is configured to use a carrier sense multiple access (CSMA) medium access control (MAC) protocol for transmission through the shared wireless medium.

We investigate CTP performance under several conditions, summarized in Table 1. The experiments are performed in a covered parking area with little foot traffic and several cars parked nearby, an open parking area, and a local road, highway, or an urban area. The condition of an experiment can be either *static* or *dynamic*, with the latter referring to cases when the engine is on, or passengers move in and out of the vehicle, or WiFi interference is present.

Sensors Data Collection

In real deployment of intra-car WSNs, data packets contain physical measurements of sensors such as tire pressure in a tire pressure monitoring system (TPMS), engine torque, or transmission pressure. For instance, a transmission control module (TCM) controls automatic transmissions in modern vehicles. TCM uses readings from transmission sensors as well as data provided by the ECUs to change gears for optimum performance in terms of fuel efficiency and shift quality [12]. In the experiments, however, we have abstracted the payload of packets (Fig. 2) into information that is needed to calculate performance metrics of the network. For instance, when a sensor node receives a packet, it updates the payload of the packet with some metadata (e.g., hop ID and RSSI information) and forwards it to the next hop. Payload fields are then processed in a post-experiment step using our developed toolkit in [13] to calculate the performance metrics and display the evolution of the network topology. In the experiments, sensor nodes generate packets with a size of 32 bytes including 8 bytes of CTP data packet header, 12 bytes of payload, and 12 bytes of CC2420 header.

Throughout the experiments, we measure the average delivery rate, delay, and number of transmissions per packet, denoted by Tx count. Assuming that in an experiment lasting T seconds, sensors generate M packets in total of which N packets are successfully received by the root. Among the total N received packets, N_u packets are unique (i.e., all duplicates are removed from the count). The delivery rate is then defined as the ratio of uniquely received packets to the total number of generated packets ($(N_u/M) \times 100$ percent). The delay of a delivered packet is the time elapsing from its generation at the source node until its reception at the root. Tx count is the number of total transmissions (including at relay nodes) per received packet at the root.

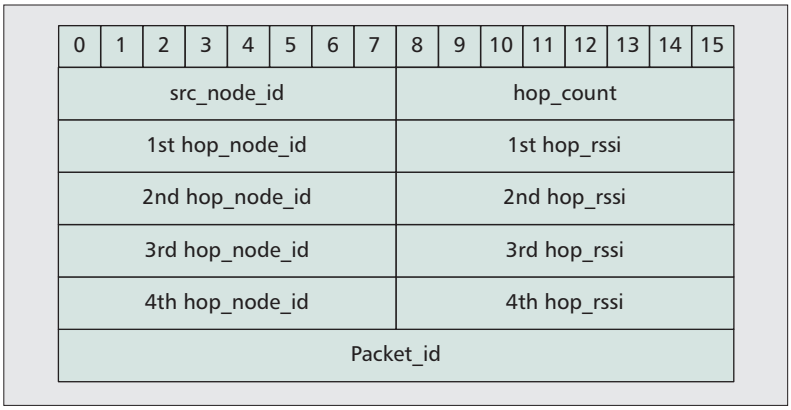


Figure 2. Payload fields of a packet.

Location	Scenario	Dynamic conditions		
		Engine	Passengers	WiFi
Parking				
Covered area	Static	Off	No	Weak[1]
	Engine-on	On	No	Weak
	Passengers	Off	Yes	Weak
	WiFi interference	Off	No	Strong[2]
Open area	Static	Off	No	Weak
Driving				
On the road	Bumpy road	On	Yes	Weak
	Highway	On	Yes	Weak
	Urban area	On	Yes	Strong[3]

[1] In-band WiFi interference is negligible.
[2] Controlled WiFi interference is exerted to the experiment setup.
[3] There is considerable urban wireless interference.

Table 1. Experimental scenarios.

Experimental Results: Static Conditions

In this section, we compare the performance of CTP with a single-hop star protocol under static conditions.

Reliability

In a star protocol all nodes communicate directly with the root, and thus, if some node-to-root links experience deep channel fading, the quality of service for that node degrades, while an intra-car WSN should guarantee high reliability across all nodes. For instance, in a TPMS network, each sensor node attached to a tire should be able to successfully deliver measured parameters to the central processors. Within this context, we examine the delivery rate of individual nodes using CTP and the star protocol. As the results in Fig. 3 show, CTP provides a high delivery rate across all nodes, whereas the performance of the star protocol varies significantly. For example, node 4 has a delivery rate of 78 percent with high variance, showing that nodes can have varying performance over time depending on the link conditions. From the network topology shown in Fig. 4, we observe that the topology induced by CTP is indeed multihop, in counterpoint to the single-hop topology of the star protocol.

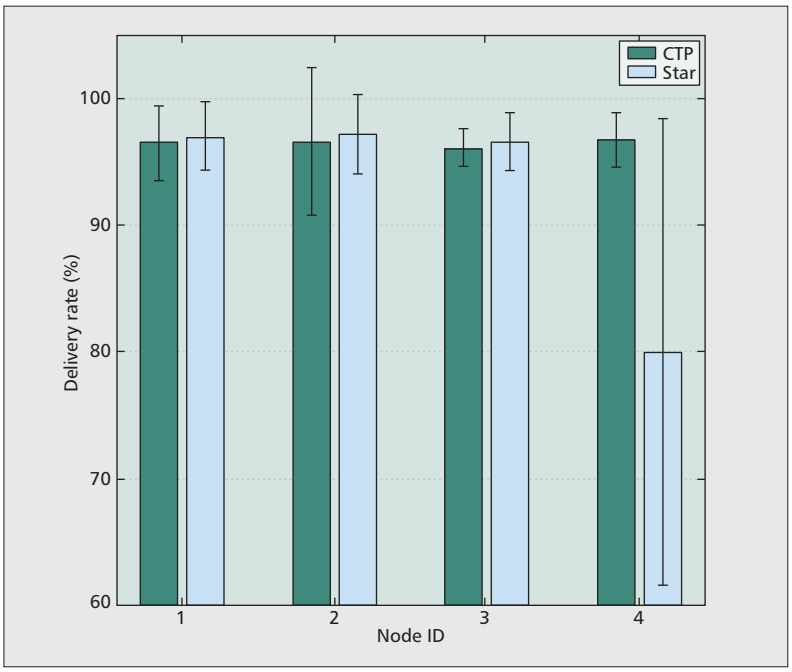


Figure 3. Delivery rate of CTP vs. the star protocol under static conditions, Tx power –20 dBm and generation rate 15 pkts/s per node. The error bars show one standard deviation from the mean over 5 six-min experiments.

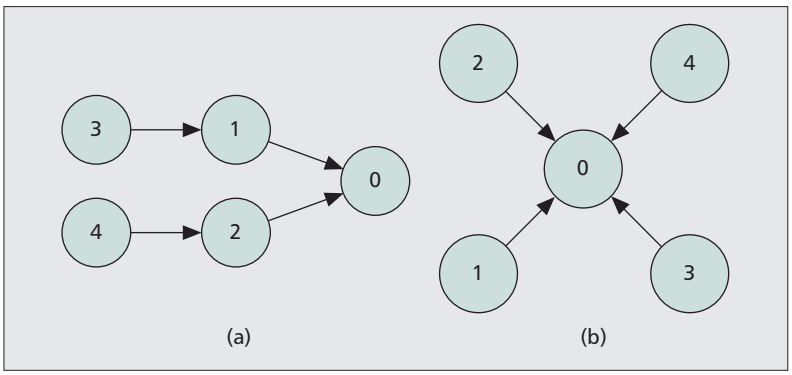


Figure 4. CTP multihop topology vs. the single-hop topology of the star protocol: a) CTP topology; b) star topology.

Condition	Tx power	Protocol	Tx count	Delay (ms)	Delivery
Static	–20 dBm	Star	1.14	14.36	92.69%
		CTP	1.72	37.35	96.43%
Passengers	–10 dBm	Star	1.22	16.72	92.59%
		CTP	2.54	82.92	98.41%

Table 2. CTP and the star protocol performance. Tx count and delay are calculated based on the *delivered* packets.

Communication and Latency Trade-offs

In wireless sensor networks power consumption and delay have significance for battery-powered and delay-limited applications.

While comparing multihop with single-hop, some trade-offs emerge: a multihop topology may require more transmissions per packet due to relay nodes, but nodes can transmit with lower power to achieve the same delivery rate as a single-hop protocol. For instance, our experimental results show that CTP with a transmit power of –10 dBm can provide the same reliability as the star protocol with a transmit power of 0 dBm, at a cost of only up to 69 percent higher Tx count per packet. Specifically, the radio of TelosB motes (CC2420) draws 11 mA current to transmit at a power of –10 dBm and 17.4 mA to transmit at a power of 0 dBm [14]. On the other hand, the receive mode requires a current of 18.8 mA, noting that in our application, the radio is almost always listening for incoming messages, which implies that the radio consumes roughly the same amount of receive power regardless of transmission activity. Based on the number of transmissions and transmit power, we conclude that CTP provides energy savings for the radio component. Previous experiments and simulations in [15] show that energy consumption by the radio is dominant in wireless sensor motes. However, it should be noted that other components, such as the processor, consume energy, and that their energy consumption depends on the instructions run by the mote.

Single-hop delay is also expected to be smaller than multihop delay, but a single-hop protocol is not always reliable, and retransmissions are required. Table 2 compares the average Tx count and delay performance of CTP and the star protocol. The results confirm that a star protocol requires fewer transmissions and incurs lower delay, but CTP provides a higher delivery rate. While the average delivery rate of the star protocol may not be significantly worse than CTP, individual nodes can experience a low delivery rate within the star protocol (e.g., node 4 in Fig. 3).

Reliability in Larger Networks

To generalize the previous reliability results, we merge the suspension network and engine network into one network with eight sensors and one collection root (Fig. 1). From the delivery rate results shown in Fig. 5, one can notice that

the performance of CTP is more reliable and stable across all nodes. In fact, different nodes in the star protocol achieve varying levels of reliability, but CTP reduces variance in performance among the different nodes, and all of them achieve a delivery rate higher than 98 percent.

Experimental Results: Dynamic Conditions

We next investigate the performance of CTP and the star protocol under dynamic in-vehicle conditions.

CTP vs. Star Protocol under Dynamic Conditions

An intra-car WSN can experience various environmental conditions, such as when the car travels through areas with intense wireless interference or passengers are sitting inside the vehicle, which can cause link fluctuations. However, an intra-car WSN should provide robust performance regardless of environmental conditions. Within this context, we compare the performance of CTP and the star protocol under two common dynamic scenarios:

- When passengers are sitting inside the vehicle
- In the presence of WiFi interference

Figure 6a shows the delivery rate of individual nodes when two passengers move in and out of the vehicle. The results indicate that CTP provides stable and reliable performance across all nodes, whereas the star protocol has a worse delivery rate. For instance, only 82 percent of the packets generated at node 2 are successfully received by the root. Figure 6b compares the performance of CTP and the star protocol in the presence of WiFi interference. Likewise, we observe that there are individual nodes within the star protocol that fail to achieve high delivery rate. The network topology induced by CTP under dynamic conditions is shown in Fig. 7, which confirms that multihop topology can provide robust performance under dynamic conditions.

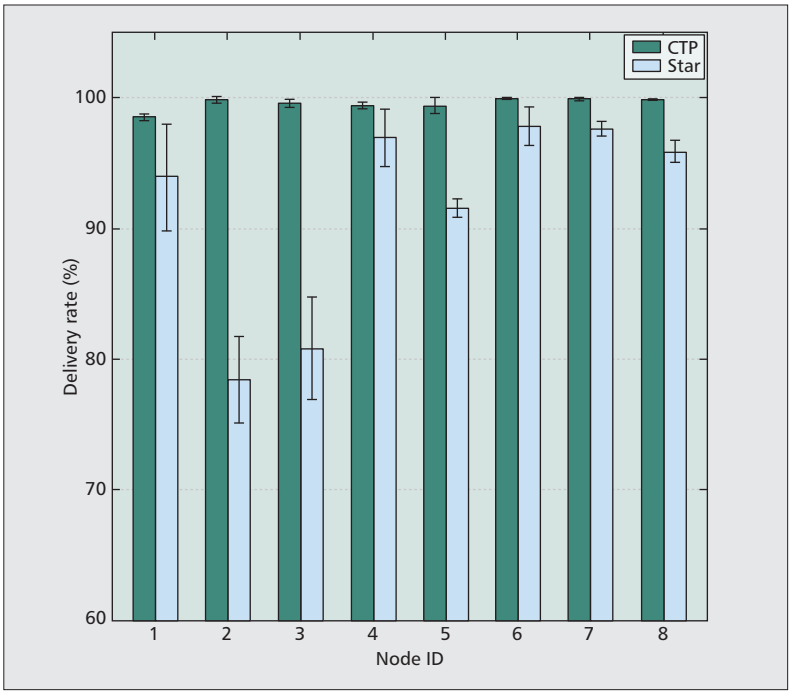


Figure 5. Delivery rate of CTP vs. the star protocol under static conditions for the eight-node network. The packet generation rate is 5 pkts/s per node, and the transmission power is set to –10 dBm.

CTP Supplementary Experiments

In this part, we extend the experiments of CTP to investigate its performance sensitivity to various environmental conditions.

Open Area vs. Covered Area — Due to the broadcast nature of wireless signals, the presence of surrounding objects, like walls and ceilings, affects wireless network performance. Within

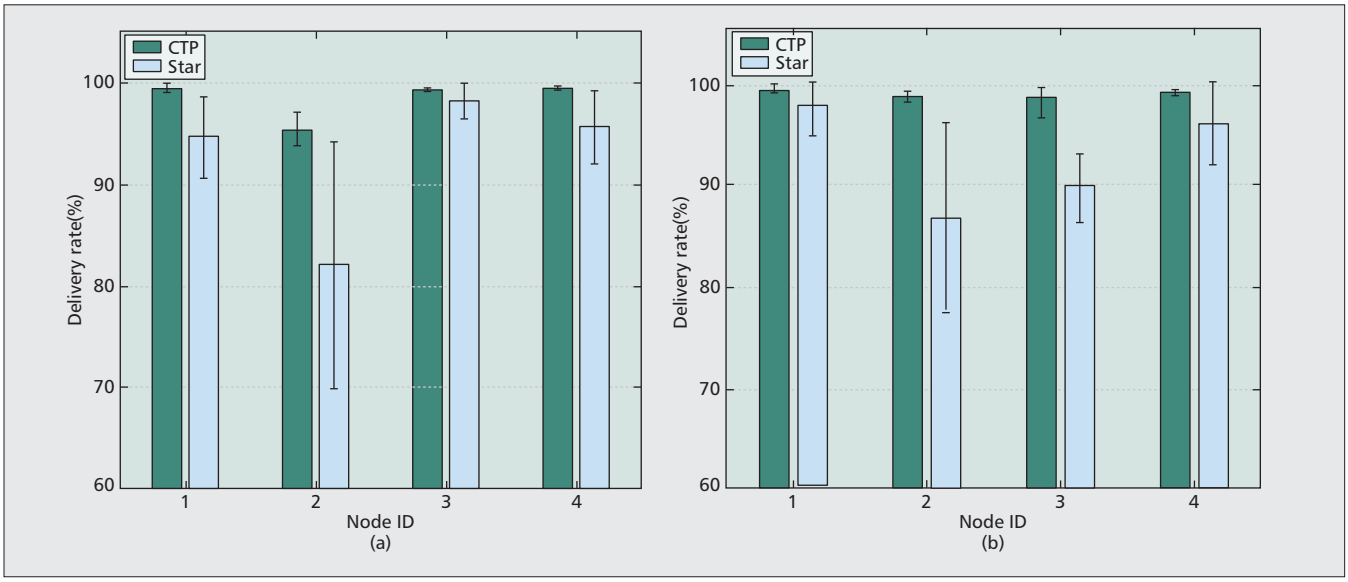


Figure 6. Delivery rate of CTP vs. the star protocol (a) under passengers-move-in-and-out scenario (b) withWiFi interferences, Tx power –10 dBm and generation rate 10 pkts/s per node. The error bars show one standard deviation from the mean over 5 six-minutes experiments.

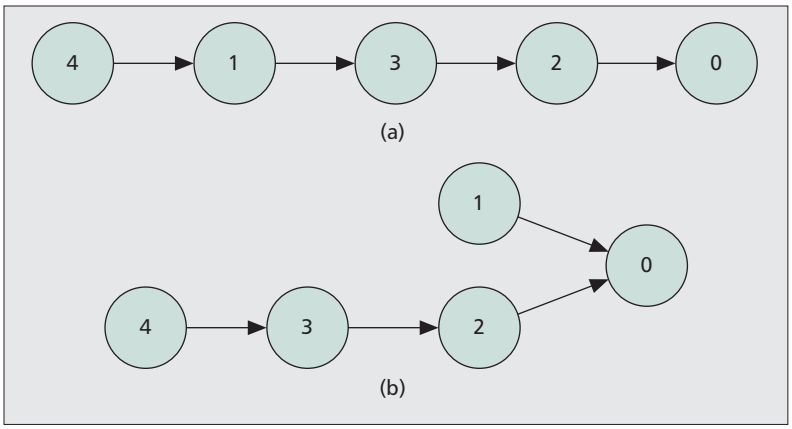


Figure 7. Multihop CTP topology under a) passengers-move-in-and-out scenario; b) with WiFi interference.

this context, CTP performance is examined in both covered and open area parking. Results shown in Fig. 8 indicate that CTP achieves better performance in the covered area parking. We speculate that this could be due to more multipath signaling in the covered area, which adds up to create a strong, almost constant signal at the receiver; an open area is potentially a poor multi-path environment, which is consistent with the experimental results in [10].

Engine-On, With Passengers — Physical channels inside a car can be highly dynamic due to external disturbances such as when the car engine is on, or when passengers move in and out of the vehicle. Figure 9 shows the performance of the *suspension network* under such conditions. The results illustrate that CTP achieves a delivery rate above 98 percent, and that engine noise does not have considerable effect on overall system performance. However, delay and Tx count performance degrade with passengers' movements, indicating that the human body can cause channel fading which may be attributed to the large attenuation of biological tissues in the 2.4 GHz range [16].

Driving Experiments — Data collection from intra-car wireless sensors is complicated by the variety of conditions a car experiences. As such, driving experiments play a critical role in performance assessment of an intra-car WSN. To fulfill this evaluation goal, CTP experiments are performed within three driving conditions: a bumpy road with poor road quality, a highway with sparse and high-speed vehicular traffic, and an urban area with dense vehicular traffic and wireless interference (WiFi, Bluetooth, etc.). Driving results shown in Fig. 10 illustrate that both networks have a delivery rate higher than 90 percent under various driving conditions. The performance of the *suspension network*, the sensor nodes of which are attached to the suspension system of wheels, is most affected when driving on the bumpy road. High-speed driving on a highway does not have a noticeable effect on both networks, and driving in urban areas does not have considerable effect on the suspension network either, but the delivery rate of the engine network degrades in urban areas, presumably due to wireless interference.

WiFi Interference — Sensor radios based on the IEEE 802.15.4 standard can suffer greatly from external interference, as they use the same frequency band as the IEEE 802.11 standard. To investigate the performance of CTP under inter-

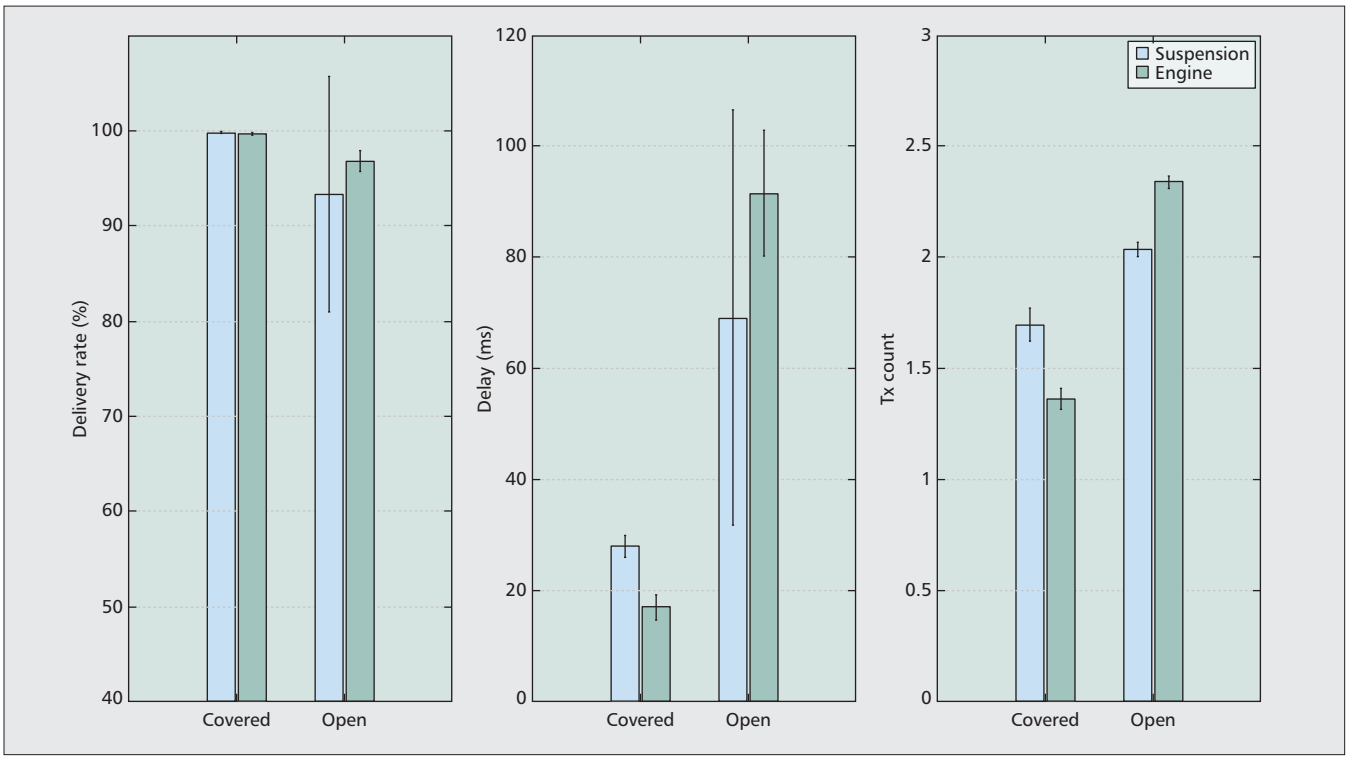


Figure 8. CTP performance of the *suspension network* and *engine network* in covered and open area parking with Tx power –10 dBm and packet generation rate 10 pkts/s per node.

ference, we run a series of controlled WiFi experiments wherein sensor nodes are configured to operate on the same frequency as a local WiFi network (channel 22 under 802.15.4 and channel 11 of 802.11). External interference is applied within two intensity levels: *light WiFi*, which reflects "normal" WiFi usage by local users, and *heavy WiFi*, which is exerted by streaming a video on a close proximity computer located in the passenger's seat. From the results shown in Fig. 11, we observe that the engine network is more vulnerable to WiFi interference, especially with low transmission power. This observation is consistent with driving experiment results in the urban area with wireless interference. It is also evident that intense WiFi interference deteriorates both networks' performance due to low signal-to-noise ratio at the nodes' receivers.

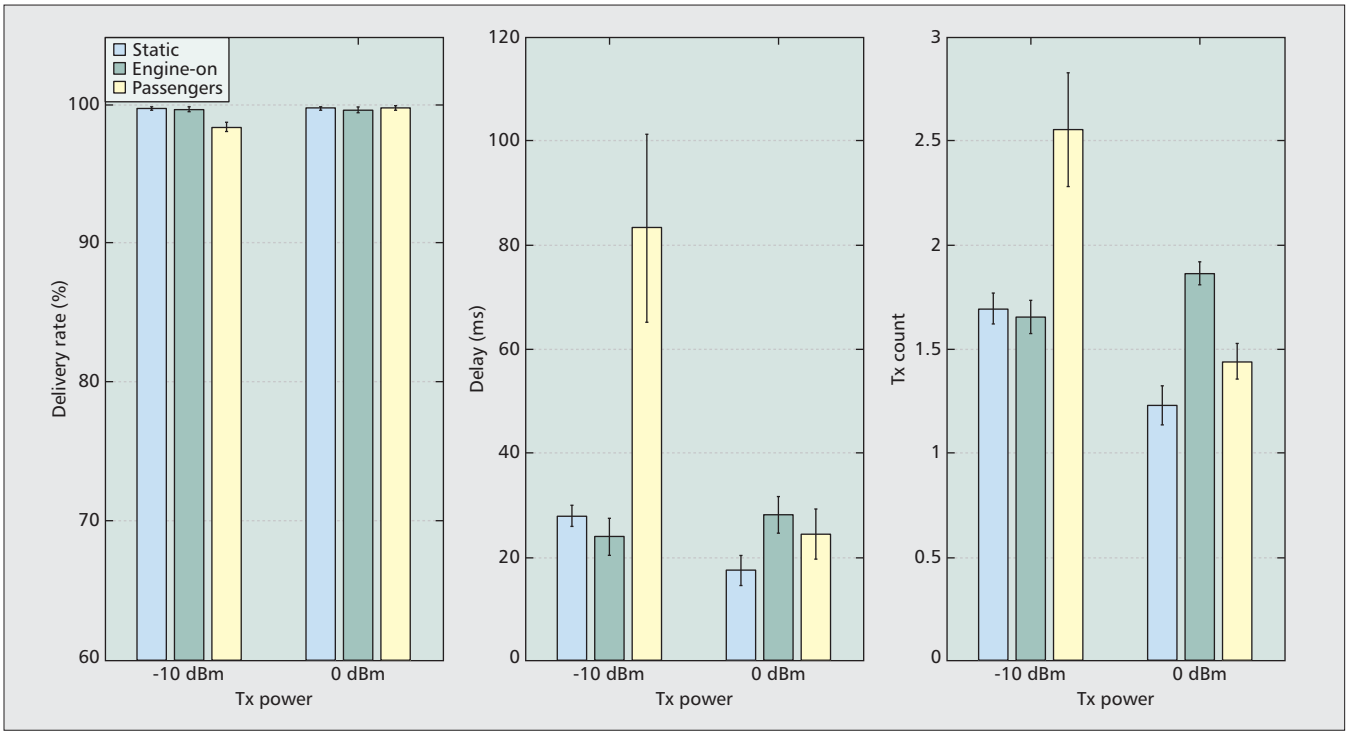


Figure 9. CTP performance under engine-on and passengers-move-in-and-out conditions for the suspension network and a packet generation rate of 10 pkts/s per node.

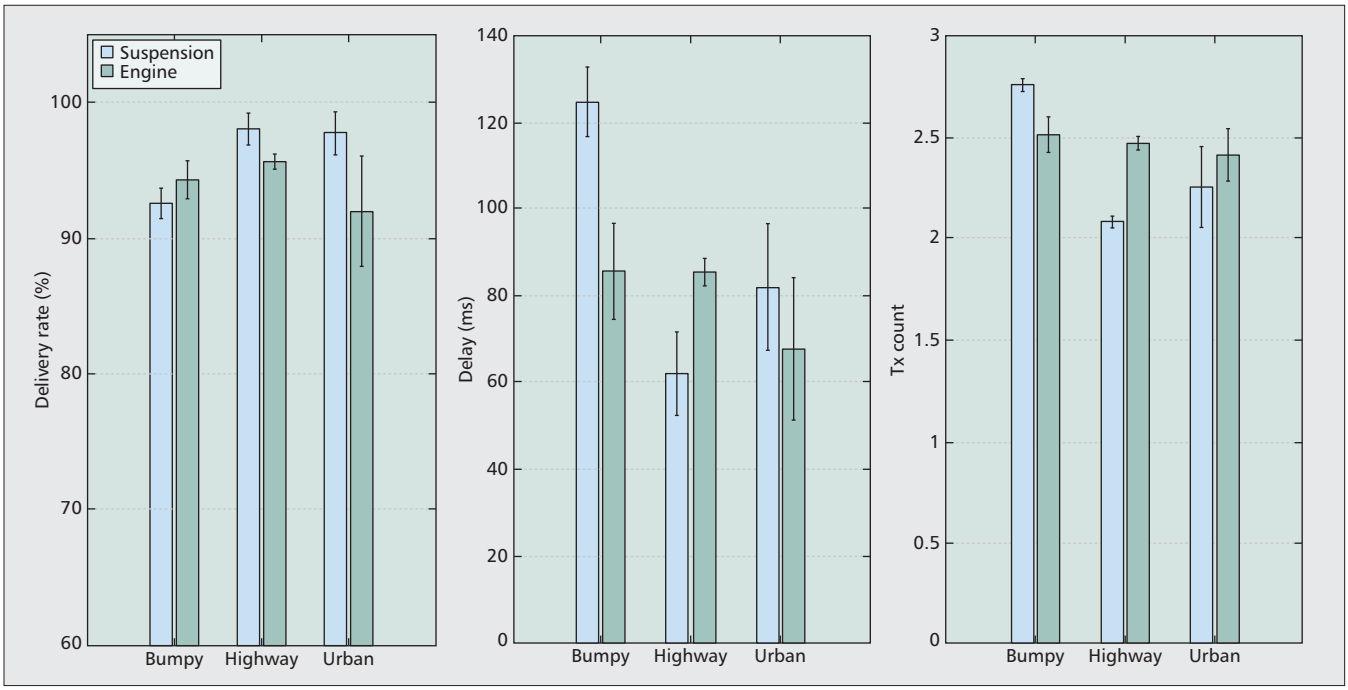


Figure 10. CTP performance of the suspension network and engine network within various driving conditions. Tx power is set to −10 dBm, and the packet generation rate is 10 pkts/s per node.

Conclusion

In this article we have investigated the performance of CTP as a multihop data collection approach, as a counterpoint to the star protocol that dominates the existing literature for intracar wireless sensor networks. Through experiments, we demonstrate that the delivery rate of the star protocol can be low in practical scenarios, while CTP achieves a reliability of more than 90 percent across *all nodes*. Whereas a star protocol requires fewer transmissions and incurs lower average delay than CTP, the radio energy consumption of CTP is smaller. Our experimental results indicate that environmental conditions have widely differing effects on network performance. For instance, passengers cause channel fading and degrade overall system performance,

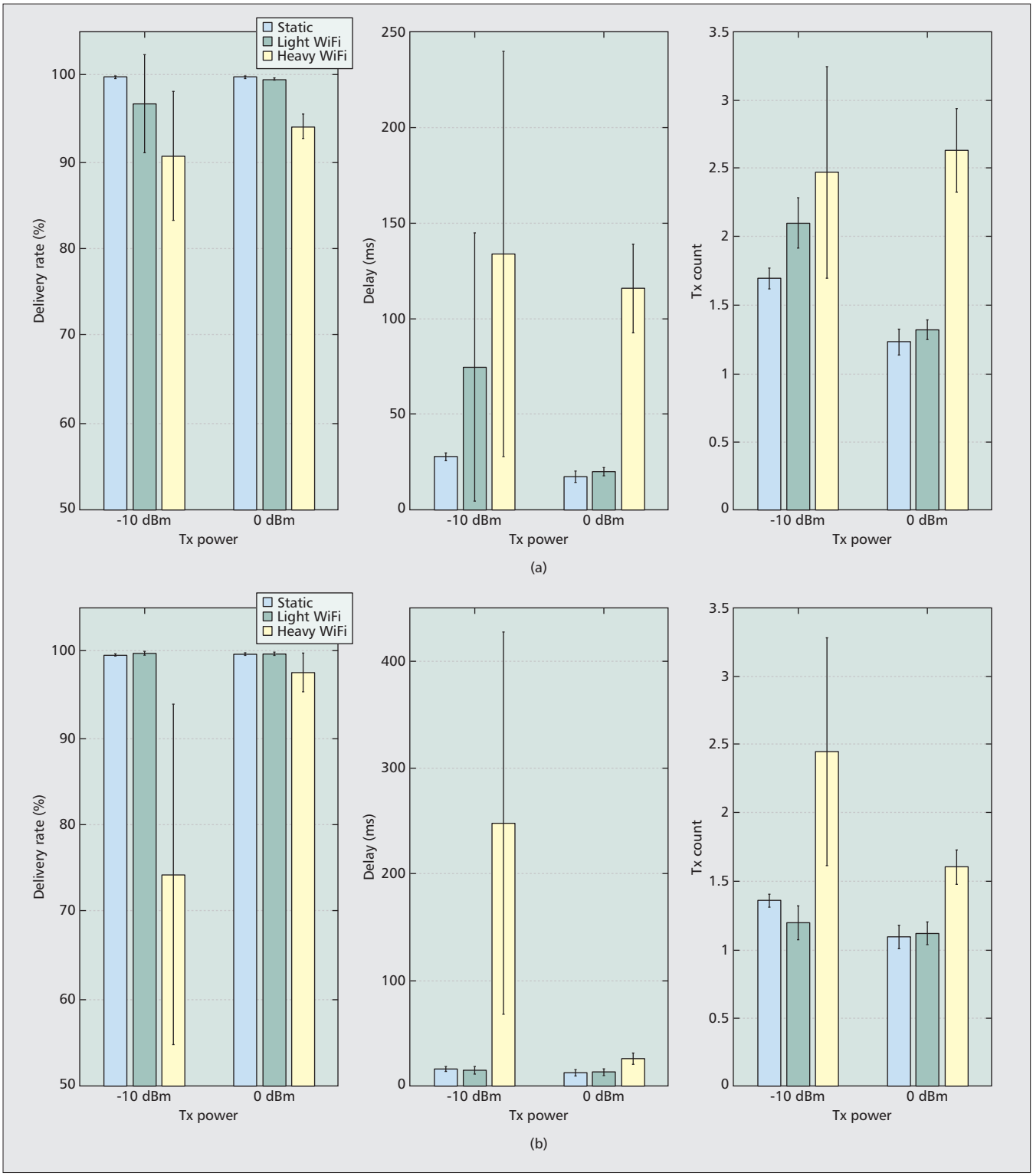


Figure 11. CTP performance under external WiFi interference for a) the suspension network; b) the engine network with packet generation rate 10 pkts/s per node.

while engine noise on the order of 2–4 dB does not have noticeable effect on performance. Intense external interference (WiFi, Bluetooth, etc.) potentially deteriorate network performance, but CTP sustains its high delivery rate (higher than 90 percent) in various driving conditions. These results serve as an illustration of the cost-benefit regions of multihop WSN inside cars.

Overall, our experimental results show that multihop networking enhances many aspects of network performance, and may be suitable for intra-car networks due to their need for robust operation in harsh environments. In future work, we plan to explore hybridized collection protocols that blend reliability with other desired performance metrics such as throughput optimality.

Acknowledgment

The work presented here was supported by General Motors, Advanced Technical Center – Israel.

References

[1] CHOSeN: Cooperative hybrid objects sensor networks; http://www.chosen.eu.
[2] M. Grudén *et al.*, "Reliability Experiments for Wireless Sensor Networks in Train Environment," *Proc. IEEE Wireless Technology Conf.*, 2009, pp. 37–40.
[3] H.-M. Tsai *et al.*, "Zigbee-Based Intra-Car Wireless Sensor Networks: A Case Study," *IEEE Wireless Commun.*, vol. 14, no. 6, 2007, pp. 67–77.
[4] O. K. Tonguz *et al.*, "RFID Technology for Intra-Car Communications: A New Paradigm," *Proc. IEEE VTC-Fall*, 2006, pp. 1–6.
[5] W. Niu, J. Li, and T. Talty, "Intra-Vehicle UWB Channels in Moving and Staionary Scenarios," *Proc. IEEE MILCOM*, 2009, pp. 1–6.
[6] O. Gnawali *et al.*, "CTP: An Efficient, Robust, and Reliable Collection Tree Protocol for Wireless Sensor Networks," *ACM Trans. Sensor Networks*, vol. 10, no. 1, 2013, p. 16.
[7] T. Winter, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," 2012.
[8] M. Hashemi *et al.*, "Intra-Car Wireless Sensors Data Collection: A Multi-Hop Approach," *Proc. IEEE VTC-Spring*, 2013, pp. 1–5.
[9] H.-M. Tsai *et al.*, "Feasibility of In-Car Wireless Sensor Networks: A Statistical Evaluation," *Proc. 4th Annual IEEE Commun. Soc. Conf. Sensor, Mesh and Ad Hoc Communications and Networks*, pp. 101–11, 2007.
[10] A. R Moghim *et al.*, "Characterizing Intra-Car Wireless Channels, *IEEE Trans. Vehic. Tech.*, vol. 58, no. 9, 2009, pp. 5299–5305.
[11] MEMSIC, Telosb specification, http://www.memsic.com.
[12] K. Vaknin *et al.*, "Experimenting with a Wireless Mesh Network Towards Sensing Inside a Vehicle's Transmission, *Proc. IEEE Int'l. Conf. Microwaves, Commun., Antennas and Electronics Sys.*, 2013, pp. 1–5.
[13] W. Si *et al.*, "TeaCP: A Toolkit for Evaluation and Analysis of Collection Protocols in Wireless Sensor Networks," *Proc. IEEE Int'l. Conf. Microwaves, Commun., Antennas and Electronics Sys.*, 2013, pp. 1–5.
[14] CC2420 Radio, http://www.ti.com.cn/cn/lit/ds/symlink/cc2420.pdf.
[15] A. S. Prayati *et al.*, "A Modeling Approach on the TelosB WSN Platform Power Consumption," *J. Sys. and Software*, vol. 83, no. 8, 2010, pp. 1355–63.
[16] J. Ryckaert *et al.*, "Channel Model for Wireless Communication around Human Body," *Electronics Letters*, vol. 40, no. 9, 2004, pp. 543–44.

Biographies

Morteza Hashemi (mhashemi@bu.edu) received his B.Sc. in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2011. He is currently a Ph.D. student in electrical and computer engineering at Boston University. His research interests include error correcting code, network performance evaluation, and wireless communications.

Wei Si (weisi@bu.edu) received his B.S. degree from Shanghai Jiao Tong University, China, in 2010. Currently, he is working toward his Ph.D. degree in systems engineering at Boston University. His research interests include routing protocols for intra-car wireless sensor networks and disruption-tolerant networks, data synchronization algorithms, and queueing theory.

Moshe Laifenfeld (moshe.laifenfeld@gm.com) received his B.Sc. from the Technion in 1992, his M.Sc. from Tel-Aviv University in 1998, and his Ph.D. from Boston University in 2008, all in electrical and computer engineering. He then took a joint post-doctoral position at MIT and Boston University, where he focused on aspects of coding theory in communication networks. Since then he is with General Motors, focusing on R&D of novel hybrid in-vehicle electrical architectures. In his past, he led the algorithms development of a third generation UMTS transceiver, and held several R&D positions in medical device startups.

David Starobinski (staro@bu.edu) is a professor of electrical and computer engineering at Boston University, with a joint appointment in the Division of Systems Engineering. He received his Ph.D. in electrical engineering from the Technion — Israel Institute of Technology in 1999. His research interests are in wireless networking, network economics, and cybersecurity.

Ari Trachtenberg (trachten@bu.edu) is a professor of electrical and computer engineering at Boston University. He received his Ph.D. and M.S. degrees in computer science from the University of Illinois at Urbana/Champaign in 2000 and 1996, respectively, and his S.B. from MIT in 1994 in math with CS. His research interests include cybersecurity (smartphones, cryptography), networking (security, sensors, localization), algorithms (data synchronization, file edits, file sharing), and error-correcting codes (rateless coding, feedback).

External intense interference (WiFi, Bluetooth, etc.) potentially deteriorate network performance, but CTP sustains its high delivery rate (higher than 90 percent) in various driving conditions. These results serve as an illustration of the cost-benefit regions of multihop WSNs inside cars.

Advertisers' Index

COMPANY	PAGE
Anritsu	3
IEEE Member Digital Library	25
Keysight Tech	Cover 2, 1
National Instruments	5
Rohde & Schwarz	9
Samsung	Cover 4
IEEE Sales and Marketing	Cover 3

Currently Scheduled Topics

Topic	Issue Date	Manuscript Due Date
Communications Education and Training	May 2015	January 1, 2015
Internet of Things/M2M from Research to Standards	August 2015	January 15, 2015
Software Defined 5G Networks for Anything as a Service	September 2015	January 15, 2015

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

SAMSUNG

THE NEXT BIG THING IS HERE™

Samsung GALAXY Note® 4

COMMUNICATIONS STANDARDS

A Supplement to IEEE Communications Magazine

DECEMBER 2014

www.comsoc.org

STANDARDS COLLISIONS AROUND SDN

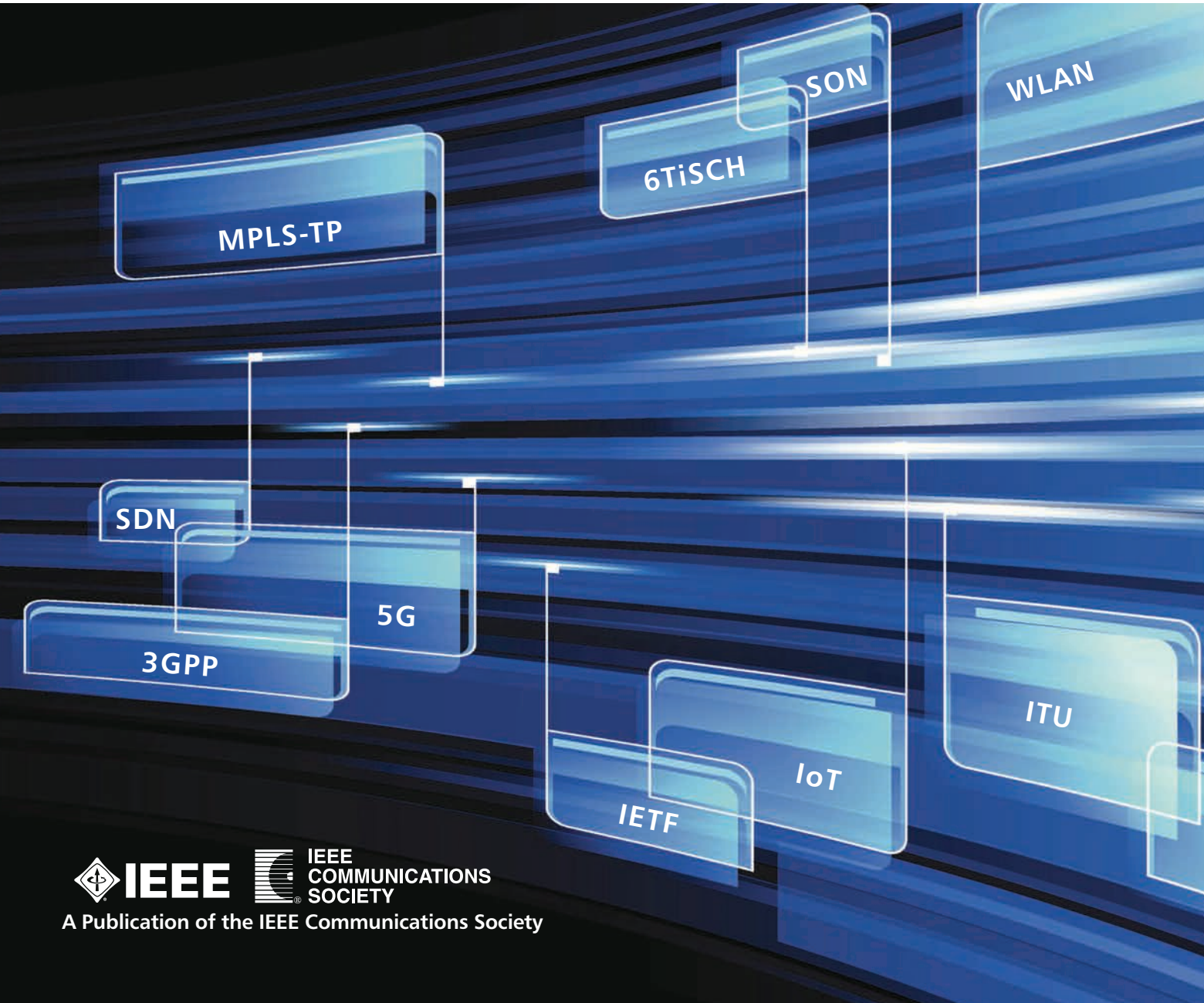
MPLS-TP LINEAR PROTECTION FOR ITU-T AND IETF

ADVACED WLAN INTEGRATION WITH THE 3GPP EVOLVED PACKET CORE

SELF-ORGANIZING NETWORKS IN 3GPP

6TISCH: DETERMINISTIC IP-ENABLED INDUSTRIAL INTERNET (OF THINGS)

5G WIRELESS ACCESS: REQUIREMENTS AND REALIZATION



IEEE COMMUNICATIONS SOCIETY

A Publication of the IEEE Communications Society

We know LTE-Advanced.

In fact, our engineers co-wrote the book on it.

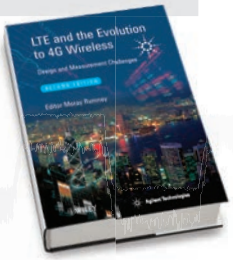
We know what it takes for your designs to meet LTE-A standards. After all, Keysight engineers have played significant roles in LTE-A and other wireless standards bodies and forums, including 3GPP. Our engineers even co-authored the first book about LTE-A design and test. In addition, we have hundreds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

HARDWARE + SOFTWARE + PEOPLE = LTE-A INSIGHTS

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

Download a free chapter of the *LTE and the Evolution to 4G Wireless* book at www.keysight.com/find/LTE-A-Insight

USA: 800 829 4444
CAN: 877 894 4414



KEYSIGHT TECHNOLOGIES

Unlocking Measurement Insights

IEEE IEEE COMMUNICATIONS SOCIETY

COMMUNICATIONS STANDARDS

A Supplement to IEEE Communications Magazine

DECEMBER 2014

SUPPLEMENT EDITOR
GLENN PARSONS

MANAGING EDITOR
JACK HOWELL

STANDARDS NEWS CONTRIBUTORS
ANDREAS KUNZ
ATHUL PRASAD
ALEX REZNIK
KONSTANTINOS SAMDANIS
JAESEUNG SONG
TARIK TALEB

2 **EXPANDING THE COVERAGE OF COMMUNICATIONS STANDARDS**
ROBERT S. FISH, VICE PRESIDENT–STANDARDS ACTIVITIES, IEEE COMMUNICATIONS SOCIETY

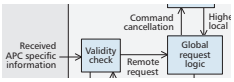
3 **LAUNCHING POINT: THE OPENSTAND VISION**
BY KAREN BARTLESON, PRESIDENT, IEEE STANDARDS ASSOCIATION

4 **COMMUNICATIONS STANDARDS NEWS**

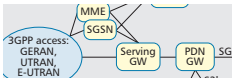
8 **EDITORIAL**

10 **STANDARDS COLLISIONS AROUND SDN**
JOEL M. HALPERN

16 **MPLS-TP LINEAR PROTECTION FOR ITU-T AND IETF**
JEONG-DONG RYOO, TAESIK CHEUNG, DANIEL KING, ADRIAN FARREL, AND HUUB VAN HELVOORT



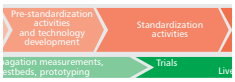
22 **ADVANCED WLAN INTEGRATION WITH THE 3GPP EVOLVED PACKET CORE**
DINAND ROELAND AND STEFAN ROMMER



28 **SELF-ORGANIZING NETWORKS IN 3GPP: STANDARDIZATION AND FUTURE TRENDS**
LJUPCO JORGUSESKI, ADRIAN PAIS, FREDRIK GUNNARSSON, ANGELO CENTONZA, AND COLIN WILLCOCK

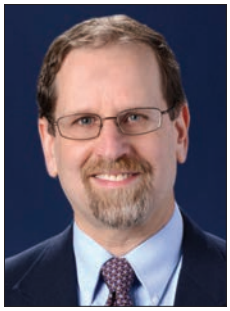
36 **6TISCH: DETERMINISTIC IP-ENABLED INDUSTRIAL INTERNET (OF THINGS)**
DIEGO DUJOVNE, THOMAS WATTEYNE, XAVIER VILAJOSANA, AND PASCAL THUBERT

42 **5G WIRELESS ACCESS: REQUIREMENTS AND REALIZATION**
ERIK DAHLMAN, GUNNAR MILDH, STEFAN PARKVALL, JANNE PEISA, JOACHIM SACHS, YNGVE SELÉN, AND JOHAN SKÖLD



EXPANDING THE COVERAGE OF COMMUNICATIONS STANDARDS

ROBERT S. FISH, VICE PRESIDENT–STANDARDS ACTIVITIES, IEEE COMMUNICATIONS SOCIETY



Robert S. Fish

As Vice President of Standards Activities of the IEEE Communications Society, I want to welcome you to the first issue of the *IEEE Communications Magazine* supplement on Communications Standards.

This supplement is the culmination of the efforts of many people to create a publication that serves the interests of the members of the global standards community who develop, use, or are otherwise interested in communications and networking standards. The aim of this publication is to cover a broad spectrum of communications and networking standards as well as standards-related disciplines, including innovation and standardization theory and methodologies, standards-related research, and standards regulations, as well as the intellectual property and socio-economic aspects of standards.

This supplement serves to expand the coverage that *IEEE Communications Magazine* has long had, predominantly through feature topic issues, on communications standards. Besides articles on standards, practical in tone, it also seeks to bring to its audience useful information about the activities of the many communication standards development groups, both inside IEEE and out.

Thanks are due to my predecessor as VP–Standards Activities, Alex Gelman, who spearheaded the effort to organize this supplement. Thanks also go to Katie Wilson, ComSoc's VP–Publications, Steve Gorshe, ComSoc's Director of Magazines, Jack Howell, ComSoc's recently retired Executive Director, and Joseph Milizzo of ComSoc's staff, for their untiring support in bringing this supplement into existence.

Finally, let me thank our inaugural editor, Glenn Parsons, for taking the reins of this endeavor and seeing that it gets off to a good start.

LAUNCHING POINT: THE OPENSTAND VISION

BY KAREN BARTLESON, PRESIDENT, IEEE STANDARDS ASSOCIATION

As I serve the remaining days of my tenure as the president of the IEEE Standards Association (IEEE-SA), I ponder the future of technology: what's been achieved today and where would this amazing world be without standards?

Thinking about life without standards is simple. The vast array of technologies that we use daily would be unable to communicate and void of quality control. The manufacturing methods alone would be disconnected and nowhere near in compliance with what consumers require. It would be chaotic and befuddling.

Standards exist for today; they exist to allow a web of electronics and mechanisms to fit like puzzle pieces in an ever-changing and complex global structure. They exist to establish trust among innovators and users of technology.

The open standards concept exists for tomorrow's technologies — advocating for a paradigm that supports principles that have helped drive the mind-blowing innovations of the past 25 years.

A borderless world — for the sake of technological innovation and open development crossing all industries — will reap major benefits for humanity, far into the future. To support this highly important initiative, IEEE-SA helped create OpenStand: a movement dedicated to promoting a proven set of principles that establish the modern paradigm for standards.

After two years of great success and achievements, IEEE-SA is celebrating OpenStand's anniversary. While in a short period of time, OpenStand's existence has invited several players to the global standards table, including W3C, the Internet Architecture Board, the Internet Society, the Internet Engineering Task Force, and IEEE — all supporting a parallel vision of market-driven standards. OpenStand has also received crucial endorsements from organizations like Adobe, Cisco, the Computer & Communications Industry Association, and others.

Collectively, they all recognize the importance of a foundational future built around open and respectful collaboration and development. The pursuit of this reality helps bring standards to life, providing multiple global organizations with an opportunity to join and contribute to this initiative.

When I think of a future with open standards, I see three core challenges on the horizon: Internet governance, privacy, and cybersecurity.

As the "connected person" and the Internet of Things become more and more interconnected in how technology functions in the world today, a free and open Internet is vital to further growth and expansion. Without it the globe could miss the critical advantages of the smart grid, augmented reality, eHealth, and many more. These technologies influence and play a crucial role in today's critical — and often controversial — issues: energy, healthcare, security, privacy, climate change, education, and communications.

Truly, to embrace a community eager to nurture the technology of tomorrow, we need to celebrate OpenStand's existence, daily.

Open standards. Open web. Open future. Where do you stand on OpenStand?

ONEM2M: THE GLOBAL STANDARDS FOR M2M AND IOT

JAESEUNG SONG
SEJONG UNIVERSITY, KOREA,
CONVENOR, TEST WG, ONEM2M

The oneM2M[1] Global Initiative [1] is a partnership project that was established in mid 2012 in order to develop specifications for a common Machine-to-Machine/Internet of Things (M2M/IoT) Service Layer. The common service layer is composed of a set of common service functions that can be easily embedded within different types of hardware and software that enable the myriad of M2M/IoT devices to connect and communicate with each other. The initiative had been started by the seven major telecommunications standards development organizations (SDOs) — ARIB, TTC (Japan); ATIS, TIA (USA); CCSA (China); ETSI (Europe); and TTA (Korea) — which are referred to as Partners Type 1. Currently, through these SDOs, more than 200 partners are registered and contributing to the development of oneM2M specifications. In addition, oneM2M has various industry member-based organizations that are Partners Type 2, including BBF (Broad Band Forum), OMA (Open Mobile Alliance), and Continua Health Alliance.

The main purpose of oneM2M is to develop standard specifications of a common M2M/IoT service layer platform that can accommodate IoT/M2M services from various vertical industries. oneM2M has developed its specifications based on a three-stage approach: stage 1 defining use cases and requirements; stage 2 designing architecture and common service functions; and stage 3 specifying protocols in details. These works have been standardized in working groups (WGs). Additionally, security, device management, and semantic enablement technologies have been standardized in order to support secure, reliable, and efficient operations of M2M/IoT services. oneM2M has defined a reference architecture that provides both basic (e.g. registration and identification) and advanced functions (e.g. location and charging). The selected architecture approach is REpresentational State Transfer (REST), which means that all things within the oneM2M system are represented by resources and managed with a set of simple commands: CREATE-DELETE-UPDATE-RETRIEVE. For protocols used between entities in the oneM2M reference architecture, onM2M reuses existing M2M/IoT related protocols such as HTTP (Hypertext Transfer Protocol), CoAP (Constrained Application Protocol — a lightweight HTTP), OMA DM (Device Management), and MQTT (Message Queue Telemetry Transport).

The first candidate release of oneM2M specifications has been delivered in August 2014 with code name "Aubergine." As shown in Table 1, the nine specifications are included in the candidate specifications in order to enable large-scale implementation of the Internet of Things. At the time this article was written, Aubergine is under technical review by various interested parties in order to improve its maturity level. Based on the feedback, Aubergine will be further revised and approved by the oneM2M Technical Plenary in January 2015. In the mean time, oneM2M has already been starting its activities such as "Home Domain Enablement" and "AllSeen and oneM2M Interworking" for the next release. With these new activities, oneM2M expects to focus on some missing features, enhance the release 1 features, and support interworking with other standards.

Spec. #	Title	WG
TS-0001	oneM2M Functional Architecture	WG2
TS-0002	Requirements	WG1
TS-0003	oneM2M Security Solutions	WG4
TS-0004	oneM2M Service Layer Protocol and API Specification	WG3
TS-0005	Management enablement (OMA)	WG5
TS-0006	Management enablement (BBF)	WG5
TS-0008	CoAP Protocol Binding	WG3
TS-0009	HTTP Protocol Binding	WG3
TS-0011	Definition and Acronyms	WG1

Table 1. List of specifications for the first candidate release of oneM2M.

REFERENCES

[1] 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.

[1] http://www.onem2m.org

WHATS HAPPENING IN THE SMALL CELL FORUM?

ALEX REZNIK
INTERDIGITAL COMMUNICATIONS, INC.,
VICE-CHAIR, SERVICES WG, SMALL CELL FORUM

The Small Cell Forum supports, promotes, and helps drive the wide-scale adoption of small cell technologies to improve coverage, capacity, and services delivered by mobile networks. The Forum has in excess of 150 members, including 68 operators representing more than three billion mobile subscribers as well as hotspot operators, telecoms hardware and software vendors, content providers, and innovative start-ups.

The Forum is currently working to address two very distinct areas. The first is tackling the challenges of the Rural and Remote small cell use case, culminating in Release Five, which will be launched at Mobile World Congress 2015. The Forum's view of "rural and remote" extends well beyond villages in the countryside struggling with basic cellular coverage. It also includes key remote deployments such as offshore oilrigs, mines, and disaster recovery zones where previous infrastructure has been destroyed. From the backhaul through to the business case and RF environment, these scenarios present a new set of challenges for both operators and vendors, and it is these areas that will form the basis of Release Five.

The second area of focus for the Forum is Virtualization. Our approach to this broad topic follows two tracks. The first looks at ETSI NFV use case #5 (virtualized 3GPP core) which is about applying NFV techniques to 3GPP core network elements. An early work item will form part of Release 5, where we will concentrate on how virtualization of the core network elements associated with the small cell architecture can lower the barriers to entry for small cell adoption. So instead of requiring dedicated hardware that has been scaled to support Tier 1 residential and enterprise requirements for millions of small cells, NFV can be used to "scale down" the implementation, leveraging standard compute, storage,

and networking components to now enable small cell solutions for a wider set of deployment scenarios. This is very interesting to the Rural and Remote market segment, but more generically should provide a set of tools to lower barriers to entry for all small cell market segments.

The second track is examining the applicability of ETSI NFV use case #6 (virtualized base station) to the small cell architecture. By examining this area we want to understand how the decomposition of the small cell base station will allow enhanced radio capabilities to be realized. This could then enable enhanced coverage/capacity or interference co-ordination, with virtualized base station functions providing control and visibility over a number of remote small cells.

Another important area of long-term focus for the Forum has been enabling application developers to take advantage of the unique user proximity and context characteristics that small cells can deliver as application presence points. We have developed a rich set of Service APIs, based on OMA's NetApi, and are now in the process of making these available to developers via Open Source licensing. The first release, including our Zonal Presence and Location Streaming API, is expected to be available for Mobile World Congress 2015.

The Forum has seen much success with the Release Program to date, and it is clear that the entire industry, from operators down to chipset vendors, appreciate the work that we do. The work taking place for both Release Five and in addressing virtualization is being driven by operator requirements, and they are both areas with much potential for the furthering the mobile industry. For more information go to http://www.smallcellforum.org/contact-contact-us

3GPP System Architecture News

Andreas Kunz
NEC Europe Ltd, Germany

The 3GPP System Architecture working group (SA2) is working in its current Release 13 on a widespread field of topics such as mission critical communication for public safety applications, system enhancements for machine type communication, user plane congestion management, IP flow mobility for Wireless LAN interworking, as well as new concepts like the dedicated core networks for specific services or IP Multimedia Subsystem (IMS) related work on webRTC interworking and Voice over LTE paging policy differentiation.

As a major topic, the work concerning public safety will be highlighted here in more detail. In Release 12, 3GPP already specified the Group Communication System Enablers for LTE (GCSE), i.e. with the help of the Multimedia Broadcast/Multicast Service (MBMS) it is possible to send messages/voice/video to a specific group in a localized geographic area. This includes traditional unicast transmission to group members not located in this area. The Proximity Services (ProSe) feature was standardized in Release 12 under the umbrella of the public safety work, but also contains commercial aspects that define the capability of mobile devices (in 3GPP called UE for User Equipment) to discover each other either directly with a special LTE radio interface enhancement or via location tracking in the core network. Once the discovery was successful it is possible to exchange information directly between those devices. Public safety enhancements for ProSe allow in addition the support of one-to-many communication, direct communication when the device is out of coverage of a macro cell, and UE-to-Network relays (now handled in Release 13). The UE-to-Network relay itself is under macro cell coverage and connects out of coverage devices via the direct communication radio interface to the macro cell. The ProSe and GCSE Release 12 work items are completed and followed up in the current Release 13 with those subfeatures that could not be concluded in Release 12 due to lack of time.

Based on the strong interest in mission critical communications, the Open Mobile Alliance (OMA), ETSI TETRA and Critical Communications Evolution (TCCE) and 3GPP held a joint workshop and agreed to set up a dedicated working group with experts from all three organizations (OMA, TCCE, 3GPP) taking their related work as a basis. 3GPP has created a new working group "SA6" for all mission critical application work, effectively starting in January 2015. SA6 will closely cooperate with SA2 on all architectural stage 2 issues and for stage 3 protocol work with the Core Network and Terminals (CT) working groups.

Broadband Forum News

Konstantinos Samdanis
NEC Europe Ltd, Germany

The Broadband Forum (BBF) has recently hosted two Birds of a Feather (BoF) open sessions that stimulated discussions among all members about Network Function Virtualization (NFV) and its adoption on the Broadband Multi-Service Network. The use of Open Source software, the creation of a new abstraction layer to support NFV in Broadband Multi-Service Networks, and the migration toward NFV were considered. BBF is encouraging collaboration with the European Telecommunications Standards Institute's (ETSI) Industry Specification Group for Network Functions Virtualization (NFV ISG) in order to assure a consistent architecture to support virtualized network functions. Currently NFV is studied in the BBF by the Service Innovation Marketing Requirements (SIMR) Group, which conducts a business analysis considering different use cases, while the End-to-End Architecture Group is exploring the migration toward the NFV paradigm after finalizing their work on multi-service broadband network architecture and nodal requirements (TR-178) and on policy convergence for next generation fixed and 3GPP wireless networks (TR-300).

The SIMR Group also studies the adoption of Software Defined Networks and Flexible Service Chaining in broadband multi-service networks analyzing the business potential and high-level requirements considering a number of use cases, while a new work item that has received attention lately is Fixed Access Sharing that deals with different paradigms on passive and active sharing of equipment and network capacity. The IP/MPLS and Core Working Group is mainly active in optimizing IP over Dense Wavelength Division Multiplexing (DWDM) interfaces, while it recently published two technical documents, one on Energy Efficient Mobile Backhaul (TR-293) that provides energy savings on Ethernet, optical, and small cell environments, and another on MPLS in Carrier Ethernet Networks (TR-224) that provides architecture and requirements for implementing MEF Ethernet service with an MPLS network. A new work item is also initiated to provide architecture and equipment requirements implementing MEF Ethernet services using BGP MPLS based Ethernet VPNs (EVPN) in IP/MPLS networks.

BroadbandHome efforts concentrate on adding more tests to TR-069 through their work on enabling network throughput performance tests and statistical monitoring, which facilitates time-based testing and diagnostic testing. The Operations and Network Management as well as the Fiber Access Network Working Groups are progress-

ing Fiber to the Distribution Point solutions, looking to adopt the NETCONF YANG management model, while the metallic Transmission Working Group focuses on a number of test suits, including the G.fast certification test plan, the performance test plan for in-premises powerline communications systems, and testing of G.993.5 self-FEXT cancellation (Vectoring).

REPORT FROM THE 3RD GENERATION PARTNERSHIP PROJECT (3GPP) – RADIO ACCESS NETWORK WORKING GROUPS

ATHUL PRASAD
NOKIA TECHNOLOGIES, ESPOO, FINLAND

The 3rd Generation Partnership Project (3GPP) has specified a 4th generation cellular technology denoted as 3GPP Long-Term Evolution (LTE) providing high speed wireless communications, for commercial and public safety services. The specification of the first version of 3GPP LTE technology was completed in 2008. 3GPP continued to develop the technology further since then, with the latest version of LTE coming to a conclusion in March 2015, when the Release 12 specifications are expected to be finalized and frozen.

In Release 12, significant attention has been given to further developing small cell deployments, including dedicated small-cell carriers. In such scenarios, the introduction of 256QAM downlink modulation will increase the peak data rates, enhancements to radio-interface based synchronization of small cells will simplify their deployments, and enhancements to discovery of small cells will allow them to be temporarily in 'off' state to reduce the created interference and power consumption, while still remaining visible to users in their vicinity. Furthermore, the concept of 'dual connectivity' has been introduced, in which the device can be connected simultaneously to a macro cell and a small cell, under the macro cell control. This allows the network to maximize the utilization of small cells using range extension, etc., while simplifying mobility management, as the connection is controlled by the macro cell.

Direct over-the-air discovery and communications between devices is a novel feature currently under finalization for this release, which brings a significant change to LTE system architecture to enable D2D operation for Public Safety applications, including scenarios where there is no cellular coverage at all.

LTE includes both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) type of carrier operation. The Release 12 FDD/TDD carrier aggregation feature now makes it possible to serve a user jointly through the operators' FDD and TDD carriers. Other topics specified within the Release 12 timeframe include, for example, support for advanced terminal receivers based on network assistance, flexible UL/DL switching points for TDD, improved mobility in heterogeneous networks, and enablers for scheduler coordination between base stations for Coordinated MultiPoint (CoMP) operation.

Looking ahead, the standardization of Release 13 of LTE has already started. 3GPP has approved the first Release 13 LTE study and work items, including Licensed-Assisted Access (LAA) using LTE, low complexity terminals for Machine Type Communications (MTC), and Elevation Beamforming (EB)/Full-Dimension Multiple-Input Multiple-Output (FD-MIMO). The work on these topics started already in October 2014, and Release 13 specifications are expected to be finalized by mid June 2016.

In LAA the focus of the study is on a single global solution for enhancing LTE to enable licensed-assisted access using unlicensed spectrum, for example at 5 GHz frequency band(s). LAA shall utilize the LTE Carrier Aggregation feature, by aggregating LTE operation in licensed and unlicensed bands. The study will also ensure that LAA can coexist with other technologies in a fair manner, having similar impact to a WLAN network capacity as another WLAN network would.

The new low-complexity terminals for MTC are expected to support enhanced coverage and low power consumption. The aim is to further reduce cost from the Release 12 single receiver MTC terminals with solutions such as smaller supported RF bandwidth.

The study on EB/FD-MIMO for LTE is identifying schemes for extending MIMO systems to utilize elevation dimension in addition to azimuth dimension for further enhancing LTE by utilizing two-dimensional antenna array operations at the base station.

Release 13 is also expected to provide extensions to SON functionality for Active Antenna Systems, study indoor positioning enhancements for LTE and investigate many other enhancements, e.g. on the areas of carrier aggregation, small cells, and dual connectivity. The approval of additional Release 13 topics is expected to continue in the 3GPP TSG RAN#66 meeting in December 2014.

ACKNOWLEDGMENT

The author would like to thank his colleagues from Nokia Technologies and Nokia Networks for contributing to this article.

PREVIEWING THE IEEE WORKSHOP ON TELECOMMUNICATIONS STANDARDS

TARIK TALEB
FOUNDER AND GENERAL CHAIR, IEEE WORKSHOP ON TELECOMMUNICATIONS STANDARDS

Building on the success of the previous editions of the "IEEE Workshop on Telecommunications Standards: from Research to Standards," this year's edition will be organized as part of IEEE Globecom 2014 in Austin, Texas, USA. This year the workshop attracted 82 submissions, with the vast majority of submissions from industry, and only 36 papers were accepted as part of the workshop proceedings. Here I would like to express my sincere gratitude to every member of the organizing committee, every member of the technical program committee, and every reviewer who has helped tremendously with the review process and the technical program creation.

The principal goal of the workshop is to close the gap between researchers, scientists, and standards experts from both academia and industry, and to promote standardization as an imperative vehicle for global consensus and cooperation among the key players, especially on the development of next-generation mobile communications systems (5G). In this regard, this year the workshop will be delivering a high quality and rich technical program, with specific focus on 5G.

The workshop will host a panel discussion on "5G: From Research to Standardization – What, How and When?" The panel will be moderated by Dr. Bernard Barani from the European Commission. The panelists are Dr. Jianzhong Zhang from Samsung Research America; Dr. David Soldani from Huawei European Research Centre; Dr. Erik Dahlman from Ericsson; and Mr. Takehiro Nakamura from NTT DOCOMO. Given the different backgrounds of the panelists, the panel will be reflecting the views of mobile operators, network equipment vendors, and user equipment vendors, as well as fund-

ing agencies, on the ongoing 5G research and standards activities. The workshop will also host two keynotes, delivered by Dr. Jochen Maes from Alcatel Lucent on "Future Trends in Hybrid Fiber-Copper Access Networks," and by Dr. Geng Wu from Intel on "Towards Future 5G Mobile Networks: From Research to Standardization to Implementation." The two keynotes will cover topics relevant to both mobile networks and fixed networks.

The workshop will also introduce a selective set of EU projects with relevance to different communications standards. A mega poster session will be organized to allow interactions among the workshop attendees.

For the first time, the workshop has recommended best paper awards considering three categories of papers: from academia, from industry, and papers that were the result of joint collaboration between industry and academia. The awarded papers for each category are, respectively:

- Xueying Song, Ge Xiaohu, Jing Zhang, and Tao Han, "An Improved IMT-A GBSM MIMO Channel Model."
- Werner Coomans, Rodrigo B. Moraes, Koen Hooghe, Alex Duque, Joe Galaro, Michael Timmers, Adriaan J. van Wijngaarden, Mamoun Guenach, and Jochen Maes, "XG-FAST: Towards 10 Gb/s Copper Access."
- Alberto Leon-Garcia and Leon Zucherman, "Generalizing MOS to Assess Technical Quality for End-to-End Telecom Session."

Given the tremendous success that the workshop has been experiencing, the workshop will be upgraded to a standalone IEEE Conference on Standards for Communications and Networking (IEEE-CSCN) starting next year. The dates and venue of the first edition of the conference are yet to be determined.

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 stake holders 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 of 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 it. This 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

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.

WELCOME TO THE INAUGURAL ISSUE

Glenn Parsons

I am honored to have been given the opportunity to initiate this supplement on Communications Standards. It is clear to me that standards enable the global market place to offer interoperable products and services at affordable cost. Standards Development Organizations (SDOs) bring together stake holders to develop consensus standards for use by a global industry. The importance of standards to the work and careers of communications practitioners motivated the creation of this publication on standards. This new quarterly publication will be incubated as a *Communication Standards Supplement* to *IEEE Communications Magazine*, which if successful, may transition into a new magazine. It is a platform for presenting and discussing standards-related topics in the areas of communications, networking, and related disciplines.

For this inaugural issue, the call for papers was necessarily broad, seeking papers related to all areas of communication and networking standards and standardization research. It is planned for future Supplements to be "anchored" around a standards topic of current market relevance to drive focus. This is similar to the feature topics in *IEEE Communications Magazine*. For example, ITU Kaleidoscope papers will be featured in 2015. Current standards topics of interest that are feature topic candidates include 5G, SDN/NFV, Internet security, and synchronization. Proposals for future standards feature topics are welcome. In this issue the reader will notice a news section from several SDOs offering current status and pointers to SDO material. In addition, for future editions we are looking to expand this with wider and perhaps more in depth columns from leading SDOs.

Despite the broad call for papers, the articles in this inaugural issue of the *Communications Standards Supplement* provide an excellent snapshot of some of the most important and topical standards technologies and issues. I trust that the reader will find these informative and illustrative of the fundamental role standards play in the communications networking ecosystem.

The first article in this inaugural issue introduces Software Defined Networks (SDN), a relatively new topic that has garnered interest in a variety of standardization bodies. While the landscape of SDN standardization can be broad, the simultaneous efforts in bodies like ONF, IETF, ETSI, and others result in duplicate work where there could be reuse. Halpern thoroughly reviews the current SDN activities and indicates that there is room for effective collaboration, based on mutual respect for work done in many places. This is easiest with a clear demarcation between standardization bodies, aided by common participants actively working in multiple groups. Finally, he highlights that the emergence of open source software as an important factor in the standardization landscape poses challenges.

The second article by Ryoo *et al.* provides a tutorial on the standardization of MPLS-TP linear protection motivated by the previous lack of agreement on MPLS-TP OAM. The background interaction between ITU-T and IETF on MPLS-TP OAM resulted in two incompatible standards being approved. Linear protection switching is critical to the development of packet switching transport networks based on MPLS technology. The article describes the options considered and the evolutionary approach adopted for the MPLS-TP linear protection specifications. A single standard, published by both the IETF (the home of MPLS) and the ITU-T (the home of transport networks) makes this situation considerably simpler compared with the case of OAM in MPLS-TP networks.

The article by Roeland *et al.* indicates that the huge growth in mobile broadband traffic in the last decade has motivated the inclusion of WLAN technology in the 3GPP Evolved Packet Core (EPC). 3GPP wireless technologies (e.g. UMTS, LTE) can be managed by operators, while IEEE WLAN technologies are not integrated. As a result, interworking or handover are not possible, resulting in an inconsistent end-user experience. This article describes the background and rationale for integrating WLAN with 3GPP EPC. A first phase of standardization, already in place at 3GPP, is summarized in detail. It defines an architecture for WLAN integration that allows a mobile device to set up a WLAN IP connection that gets routed to EPC. The article also explains a potential second phase of WLAN integration with EPC, which covers more advanced WLAN features including handover with IP address preservation, attachment to a non-default APN, EPC offloading, and support for multiple IP connections over WLAN.

Following on with the 3GPP standards activities, Pais *et al.* describe how self-organizing networks (SON) provide mobile operators the ability to provide better performing services at reduced costs. The article gives an introduction to mobile network automation and describes SON functions that have already been standardized in 3GPP. Previously these network operations were mainly slowly changing network configuration, but the trend toward automated and dynamic network operations demands advanced SON features. The article describes the existing SON functions standardized up to 3GPP Release 11 and hints at enhancements planned for 3GPP Release 12. The article concludes with the reality that despite the benefits of SON standards, several deployment challenges remain for mobile operators to successfully incorporate SON functionalities.

While underlying wireless standards are necessary, an effective integration with the Internet Protocol (IP) is necessary to realize the evolution toward an Internet of things. The integration of low-power wireless networks into a larger networking system in the industrial vertical is the focus of the article by Dujovne *et al*. These low-power wireless technologies based on Time Synchronized Channel Hopping (TSCH) satisfy the stringent reliability and low-power requirements of industrial applications, and are fundamental to standards such as WirlessHART, ISA100.11a, and IEEE 802.15.4e TSCH. IETF standards like 6LoWPAN allow low-power wireless devices to behave like any other Internet host. Still, the article highlights that further standardization is needed to allow TSCH schedules to be managed in an IP-enabled infrastructure, those empowering industrial performance with the ease-of-use of IP. This is the goal of the newly created IETF 6TiSCH working group ("IPv6 over the TSCH mode of IEEE 802.15.4e").

The final article illustrates how standardization of globally relevant and anticipated technology must have a plan. Selen *et al.* look to the future and what 5G might look like, but are clear in that there is a standardization timeline already in place at ITU-R for 5G. 5G is the wireless access technology solution for the beyond 2020 time frame that will provide access to information and sharing of data anywhere and anytime for anyone and anything — a networked society. In this paper, a view of 5G opportunities, challenges, requirements, and technical solutions is provided. This gives us a better understanding of the need for a well understood standards process to help us reach a goal that will benefit all of society.

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.

Standards Collisions Around SDN

When a technology space attracts a lot of attention, as SDN does, many standards bodies and industry associations undertake work in the area. This is good and necessary because no one body can do everything. Unfortunately, when this happens overlap and collision are common.

Joel M. Halpern

Abstract

A review of a number of ongoing SDN standardization and open-source activities, this article also discusses the interactions both actual and potential in various standards bodies.

Introduction

This article is an initial examination of the ongoing collisions among disparate standards bodies and open-source communities regarding the future of Software Defined Networking (SDN) standardization. When a technology space attracts a lot of attention, as SDN does, many standards bodies and industry associations undertake work in the area. This is good and necessary because no one body can do everything. Unfortunately, when this happens overlap and collision are common. At best such collisions lead to duplicative work and, frequently, to incompatible standards.

Defining SDN

When technologies are new and interesting, technology developers frequently want to associate their activities with that interest. Today this is happening with SDN. As a result many technology and product announcements are positioning themselves as part of SDN. Thus, depending on where one looks, SDN looks more like a way of building packet-data planes, control planes, or even management systems. This complicates any discussion of SDN, particularly with respect to standardization, as participants often unknowingly discuss different technologies under the same label.

For this article we will use the definition for SDN networks published by the Open Networking Foundation (ONF) [1] that describes them as ones that are:

- Directly programmable: Network control is directly programmable because it is decoupled from forwarding functions.
- Agile: Abstracting control from forwarding lets administrators dynamically adjust network-wide traffic flow to meet changing needs.
- Centrally managed: Network intelligence is (logically) centralized in software-based SDN controllers that maintain a global view of the network, which appears to applications and policy engines as a single, logical switch.
- Programmatically configured: SDN lets network managers configure, manage, secure, and optimize network resources very quickly via dynamic, automated SDN programs, which they can write themselves because the programs do not depend on proprietary software.
- Open standards-based and vendor-neutral: When implemented through open standards, SDN simplifies network design and operation because instructions are provided by SDN controllers instead of by multiple, vendor-specific devices and protocols.

For many of the standards activities discussed in this article, this definition should be augmented explicitly with one aspect that many people assume, or even consider implicit in the above. This is that SDN often includes and discusses the separation of network-forwarding behavior from network-control behavior.

A Subset of Related SDN Standards Activities

Many standards bodies, industry fora, and open-source projects are involved in SDN activities. Many more would like to be involved. In this section, I will review four of the active and well-known bodies working in this space, and then I'll go on to discuss their interactions.

ONF

The ONF defines itself [2] as:

The Open Networking Foundation (ONF) is a user-driven organization dedicated to the promotion and adoption of Software-Defined Networking (SDN) through open standards development.

Historically, the ONF grew out of the work at Stanford University on the OpenFlow protocol. It is now an industry consortium with about 150 member companies. Work in the ONF is divided into 10 working groups. The following sections review the general aims and activities of each working group. The first working group below, Extensibility, is the most central to the efforts because it maintains the OpenFlow protocol. Next, the list progresses through the working groups from those that are most central to the work to those that are expanding the scope of the ONF in various directions.

Extensibility — Maintains the OpenFlow specification. The group has produced maintenance releases of the OpenFlow 1.0 and 1.3 protocol specifications. The group has completed the OpenFlow 1.4 specification, and has begun work on OpenFlow 1.5. All OpenFlow protocol specifications are built around the concept of match-action-tables. The protocol allows the controller to specify entries for these tables. An entry consists of actions associated with a series of matches against arbitrary subsets of the packet header fields defined in the protocol. These fields include the incoming port identifier, the Ethernet Source and Destination Address, the Ethertype, and the IPv4 and IPv6 source destination

COMMUNICATIONS STANDARDS

The author is a Distinguished Engineer with Ericsson, Inc., San Jose, California 95134 USA.

and carried protocol identifiers. They also include the port number for the protocols, such as TCP and UDP, that have those fields.

Actions are collected from the matching entry, and either applied immediately or added to a list to be applied later to the packet. These actions can perform various common packet manipulations, such as rewriting a header field or directing a packet to a port. The protocol also provides for experimental match fields and experimental actions. While the original specification supported a single table, support was added in the 1.1 release for multiple tables and context pointers to resolve the combinatorial entry explosion issues, at the cost of additional lookups.

Configuration and Management — Responsible for the protocols used to manage OpenFlow switches. This is driven by a deployment assumption that the common deployment case for OpenFlow will be a forwarding device strictly controlled via OpenFlow. Such a device still needs to be managed. This working group has produced 1.0, 1.1, and 1.1.1 versions of the specification, and is working on 1.2. The specification relies on the IETF NetConf Configuration protocol [3] for its communication mechanism. The specification documents the information directly in XML, although the work was driven from the YANG [4] work done in the IETF NetMOD working group. The version of the data model written in YANG is included as an appendix to this document.

Architecture & Framework — Describes the problem space and architectural approach to SDN and the role of the OpenFlow-related work in that space. A brief overview has been produced, and a comprehensive architectural description is approaching completion. If produced, and advanced in collaboration with other standards bodies, the resulting architecture could be very helpful to the community and community collaboration.

Forwarding Abstraction — One of the challenges with using OpenFlow is that it uses the single abstraction provided by the OpenFlow protocol for interacting with everything. The Forwarding Abstraction work, nearing publication, is intended to enable pre-runtime description of the needed forwarder behavior so as to enable more effective resource utilization.

Optical Transport — Develops specifications for using OpenFlow for control of optical transport networks. This work relies on ITU-T-developed models of optical transport networks to define the relevant components. It also differentiates between the control of detailed optical properties and the control of traffic flow and optical path placement.

Northbound Interface — Recently began working on defining the interfaces an OpenFlow-based SDN controller would expose to other policy and control elements in the network, particularly those operating at a higher level of abstraction. The work is in the very early stages.

Wireless and Mobile — Collects use cases, works on architecture and on protocol extensions to OpenFlow to extend the ONF-based work to wireless and mobile domains. Although the work is in the early stages, it has already divided into multiple subtasks addressing Evolved Packet Core mobile processing (EPC), Mobile Backhaul, and enterprise wireless networks.

Migration — This group is a successor to an earlier ONF effort to define hybrid device operation, i.e., how to structure and use a device which supports simultaneously OpenFlow and other operating paradigms. That earlier??OK? effort did not produce approved results. The migration work is aimed at the narrower but still quite significant problem of how to introduce OpenFlow into a network currently using other technologies.

Other Activities — In addition to those just listed, the ONF has several other working groups.

Testing and Interoperability is concerned with providing test cases and running interoperability events. It is also considering certification possibilities.

Marketing and Education does outreach. It produces white papers and solutions briefs, and provides speakers to talk about the ONF.

There are also several discussion groups working towards further expansion of the ONF's work. Of particular relevance to this article, the ONF has been working on defining mechanisms for service chaining, under the auspices of work on applying OpenFlow to layers 4–7.

Internet Engineering Task Force (IETF)s

The IETF provides the base standards for the Internet. It is an outgrowth of the bodies that advised the U.S. government on the original DARPANET activities. The IETF describes its goal [5] as:

> The mission of the IETF is to make the Internet work better by producing relevant, high-quality technical documents that influence the way people design, use and manage the Internet.

The IETF consists of well over 100 working groups divided into seven large areas. Most of their work is unrelated to SDN. While many of the RFCs produced by the IETF are used in networking activities (much as many IEEE standards are used in almost all networks), some specific working groups are working on SDN.

Forwarding and Control Element Separation (ForCES) — Dating to 2001, this working group was established to build upon earlier work, such as the ATM Forum Multi-Protocol Over ATM (MPOA), as well as to generalize the work then being attempted in the Network Processor Forum.

The group focuses on a semantic model for data-plane packet processing, and a protocol that allows a control element to manipulate forwarding elements by referencing elements from that model. This resulted in the publication of RFCs [6] and [7], which provide the definitions for undertaking this task. While much less well known than the OpenFlow work at the ONF, it has been used for interoperable

The IETF consists of well over 100 working groups divided into seven large areas. Most of their work is unrelated to SDN. While many of the RFCs produced by the IETF are used in networking activities, some specific working groups are working on SDN.

As a companion to work in the IETF, research is underway by the Internet Research Task Force. This group sponsors research that it is hoped will produce results useful to the Internet community but which is not yet ready for engineering. One such research group is devoted to research on SDN.

implementations [8] and I am told of corporate prototypes and commercialization efforts underway.

Interface to Routing Systems (I2RS) — A relatively recently formed working group, it was established to address a significant gap in the approach being taken to SDN. A lot of work was being done on controllers interacting with forwarding planes. However, to use SDN technologies in real networks controllers must interact with routing protocols, and SDN control must to be able to apply policy to actual routers. Those routers may be integrated devices, or may themselves be decomposed, centralized, or otherwise make use of SDN technologies. As the charter for the working group [9] says:

> "I2RS facilitates real-time or event-driven interaction with the routing system through a collection of protocol-based control or management interfaces. These allow information, policies, and operational parameters to be injected into and retrieved (as read or by notification) from the routing system while retaining data consistency and coherency across the routers and routing infrastructure, and among multiple interactions with the routing system."

Thus, the goal is to allow applications to learn from and request changes of the routing system. This leverages the power of both classic distributed routing and centralized, policy- and application-driven SDN. This working group is getting ready to publish its architecture and, as of this writing, is wrestling with the question of whether any existing protocols such as ForCES, NetConf with YANG, or RESTCong with YANG can meet its requirements.

Service Function Chaining — As discussed below under the topic of the ETSI NFV initiative, operators are driving the standards community to deliver services to customer that are more flexible, efficient and cost-effective. They see that the use of virtualization functions and traffic direction has the potential to change dramatically the way they enable and deliver services. As part of this, operators have said it is highly desirable to be able to direct subsets of traffic across the network in such a way that each virtual service platform sees only the traffic it must work with. This has been a major difficulty in managing the scaling and placement of services using conventional control mechanisms and existing networking technologies.

The Service Function Chaining working group in the IETF is developing standards for the data-plane component of service chains intended to improve this traffic-direction problem. As the charter for the working group [10] says:

> The SFC working group will document a new approach to service delivery and operation. It will produce an architecture for service function chaining that includes the protocols or protocol extensions to convey the Service Function Chain and Service Function Path information to nodes involved in the implementation of service functions and Service Function Chains, as well as mechanisms for steering traffic through service functions.

Still in its early stages, the work has seen many debates about the approaches to be taken, exactly what needs to be defined, and what room can or should be left for implementation variation. The working group is allowing for a range of carriage mechanisms (what it refers to, in an overloading of terms, as transport) so as to allow the use of Layer 2 encapsulations, such as Ethernet with VLANs, to identify service paths, intermediate technologies such as MPLS, or IP encapsulations, depending upon circumstances.

While the work does not mandate specific control mechanisms, it is widely expected that the dynamic service provisioning and the need to shift flows to different service chains as needs arise will lead to the use of SDN technologies for controlling the classification and forwarding functions in the service paths.

SDN RG — As a companion to work in the IETF, research is underway by the Internet Research Task Force (IRTF). This group sponsors research that it is hoped will produce results useful to the Internet community but which is not yet ready for engineering. One such research group is devoted to research on SDN. The group provides a forum for discussion on topics that touch on SDN engineering issues but that are more forward looking. As such, it has less interaction with the other groups covered in this article; it is mentioned here for completeness.

European Telecommunications Standards Institute Industry Specification Group for Network Function Virtualization (ETSI NFV ISG)

An operator-driven initiative, this is aimed at enabling operators to leverage newer technologies to increase their service delivery capability, drive costs down towards those of data-center networks, and increase the speed of deployment of new or experimental services. Though hosted in the ETSI, a European body, the initiative's more than 200 members come from all over the world.

As described by ETSI [11], the initiative aims to:

> … define the requirements and architecture for the virtualization of network functions and to address the technical challenges.

A clear list of the challenges being faced is provided in the same location. Strictly speaking, the focus is on virtualization. But it is well understand that the use of SDN technologies provides a powerful lever to enable many of the use cases being described. This Industry Specification Group (ISG) is driven by the network operators to meet their needs. As such, it is a critical element of the environment. It provides requirements to other activities, and provides a clear view of how the operators see the work.

The work in this body is organized into a Technical Steering group and six working groups.

Architecture of the Virtualization Infrastructure — Known as the NFV INF WG, this working group is producing a reference architecture for a virtualiza-

tion infrastructure, and the reference points for interconnecting its components. This work underpins much of the other solution-oriented effort in ETSI NFV.

Management and Orchestration — Concerned with describing the deployment, instantiation, configuration, and management of network services based on the NFV infrastructure. In particular, this group focuses on the integration between network service delivery and the operational support systems (OSS) and business support systems (BSS) central to the business processes of network operators.

Many of the use cases and problems being developed and refined in this group are aimed at service chaining and flexible service delivery. These describe work to be done in other bodies and sometimes overlaps with that work.

Software Architecture — Responsible for defining the reference software architecture of network functions to be deployed in the infrastructure and management environment defined by the respective working groups. In particular, this work will define the detailed requirements of the interfaces and mechanisms defined by other working groups.

Reliability and Availability — Defines the reliability and availability requirements in a network function virtualization environment. Specifically, with the replacement of traditional telecommunications equipment with more data-center-oriented equipment, and with dynamic and virtualized instantiation of service functions, there are new approaches to reliability and availability. This working group will determine the impact of this evolution and how operators can meet their customers' needs in this regard.

Security Expert Group — Provides security review and advice to the broader ETSI NFV activity. The team will work across the ETSI NFV working groups, and advise the technical steering group.

Performance and Portability Expert Group — The virtualized environment dramatically changes the approach to performance and the portability requirements of the network functions being deployed. This expert group will advise the ETSI NFV working groups on performance issues, constraints, capabilities, and potential advantages of various architectural or deployment choices.

Open Daylight

An open source software activity under the auspices of the Linux foundation. As of this writing, it had 36 member companies providing resources to develop an SDN controller for a wide range of applications.

In describing the needs leading to the development of Open Daylight, the community says [12]:

> At this early stage of SDN and NFV adoption, the industry acknowledges the benefits of establishing an open, reference framework for programmability and control through an open source SDN and NFV solution. Such a framework maintains the flexibility and choice to allow organizations to deploy SDN and NFV as they please, yet still mitigates many of the risks of adopting early-stage technologies and integrating in existing infrastructure investments.

For flexibility, the software is written in JAVA. It supports a wide range of interfaces to applications, principally using REST technologies. It also includes a CLI to allow human interaction, and support for JAVA RMI for closer coupling to the software.

Similarly, the software supports a wide range of protocols for interacting with the network. These include NetConf, Simple Network management Protocol (SNMP), Open Virtual Switch Data Base (OVSDB), OpenFlow, Border Gateway Protocol (BGP), Path Computation Engine Protocol (PCEP) and Locator/Identifier Separation Protocol (LISP). The architecture also explicitly allows for developers to add other interfaces, such as for proprietary techniques to control special equipment.

The system is built around a core of YANG [4] models which describe the services, interfaces and data storage provided. This enables automatic generation of much of the code, and a common model-driven dispatch mechanism to support the flexibility needed.

For flexibility, the software is written in JAVA. It supports a wide range of interfaces to applications, principally using REST technologies. It also includes a CLI to allow human interaction, and support for JAVA RMI for closer coupling to the software.

Interaction Among Activities

It is clear from looking at the above list of activities that they require a lot of coordination. Some of this is taking place, and working well. But some is not working as well.

ETSI NFV

This group is designed to promote positive interaction with standards bodies. It has a formal collaborative relationship with the ONF to enhance SDN support of NFV needs.

Similarly, via common members, NFV has been interacting closely with the IETF. NFV requirements are driving much of the requirements work in the I2RS and SFC working groups.

ETSI is enabling and providing support for several proof-of-concept activities, many of which are expected to make use of Open Daylight software, and may also drive feature development for Open Daylight.

Each standards body is also working with ETSI as they analyze the needs and gaps in the current specifications, so that problems can be addressed promptly.

This is then an example of the best kinds of interaction. Communities recognizing needs, and working with other communities to fill those needs.

ONF and IETF Interactions

The conflict between these bodies is unfortunate. When the ONF was formed, its founders felt that due to details of the process rules and the approach to working, they could not bring their work to the IETF. Instead, it founded a new standards body, and developed a constituency focused on its specific needs. This has had some positive results; in particular, their initial specifications were developed very quickly. How-

Probably the oldest telecommunications standards body in the world, the ITU has been influential in many critical aspects of telecommunications. Currently, with respect to SDN, it is particularly active in defining architectures and requirements for the use of SDN relative to transport networks.

ever, those specifications are very narrow, with a lot of work now required to determine how to utilize them in a broader space. In addition to reducing participation in the work, this approach also resulted in a similar difficulty in allowing the IETF to use ONF products.

Also, as with many organizations, the ONF is seeking a broader role in its community, and wishes to be the central standards body for all SDN. This causes friction with other standards bodies, and complicates the interactions, which could be more effective and useful to all parties.

Such focus also leads to a common problem within standards bodies, that of seeing one's own standard as more important than the choices the market makes. An example of this is the OF-Config protocol. This protocol was developed by the ONF for managing OpenFlow switches. But the market has failed to adopt the protocol. Rather, it is using a proprietary protocol known as OVSDB. Better interaction would help mitigate the tendency to fall into this sort of pattern, as everyone gets used to using components from other places.

This collision makes development work significantly more complex. For example, the work developing the YANG models for OF-Config would have proceeded more easily if the two bodies had cooperated. Instead, a few individuals had to put in significant efforts to replace the collaboration.

Similarly, work on SDN control-plane activities in the IETF would be simpler if it were easy to collaborate with the ONF on use of the OpenFlow protocol as one of the possible control mechanisms, one that many people expect to use.

The ONF is currently engaged in developing relationships with other standards bodies and industry groups. Several of these have resulted in preliminary agreements, but there is not yet enough interaction to see if these will work well or not. It will be particularly interesting to see how the differences in rules on intellectual properties rights are dealt with as the ONF works with other bodies.

Open Daylight Interactions

This is not strictly a standards body. It is included here because its interaction with standards is so important and because it reflects an operating approach becoming more important with time.

Open Daylight includes as participants many individuals from the ONF and IETF communities. It is building software using protocols from both these standards bodies. This sort of implementation provides valuable feedback on what is clear or unclear, workable or unworkable, and useful or useless in the standards specifications. Also, in the course of implementation, the Open Daylight team may find it needs protocols or mechanisms that are not currently specified. These determinations can provide the impetus to undertake new and useful standards work.

There is one important concern to consider in these interactions. This is a tendency among participants to confuse code with a standard. Code is a way of implementing a standard. It provides a powerful reference. But being simply an implementation, code does not address many of the issues standards must deal with. It does not elucidate the core needed for interoperability, nor differentiate that from the other aspects needed for a particular implementation. Nor does it make clear where implementation choices exist. In particular, it does not indicate what the critical aspects are for implementations to interoperate. Such interoperation would lead to a robust and innovative environment that would also enhance and simplify evolution of the work. That said, the real problem with seeing the code as the standard is that this replaces a powerful collaborative relationship with a competitive one. Such a competitive view impedes the cross-fertilization needed, and harms both the standards-development and the open-source development processes.

This conflict results in difficulties particularly when the open source code is running, and the standards take a different direction. Again, individual participants aware of any trying to avoid problems do help. But more formal recognition of the mutual value of the efforts in Open Daylight and in standards-development activities would improve the situation.

Other Relevant Standards Bodies

I have just described some of the standards and industry activities in SDN, and some of the ways they may collide or cooperate. Not all the bodies mentioned are active in SDN standardization and collaboration. While no short discussion can cover all such work, the following highlights some additional activities that also overlap with work described above.

International Telecommunications Union Telecommunications Standardization Sector (ITU-T)

Probably the oldest telecommunications standards body in the world, the ITU has been influential in many critical aspects of telecommunications. Currently, with respect to SDN, it is particularly active in defining architectures and requirements for the use of SDN relative to transport networks. This can provide valuable perspective for the ongoing development of SDN protocols. Transport networks have important requirements different from those of other kinds of networks. By highlighting these needs, the ITU-T can provide valuable direction to the ongoing efforts.

Institute of Electrical and Electronic Engineers (IEEE)

IEEE supports vibrant and energetic standards activities across a broad range of problems and needs. In networking, it is particularly well-known for the IEEE 802 series of recommendations and standards. Most people simply think of this as Ethernet, but its scope is larger than that.

For example, IEEE 802.1 has been working on network virtualization (in 802.1Q) since 1998, thus providing the underlying structure and exemplar well before the term was popular.

802.1 has completed work on edge virtualization, and for virtual extensions of bridges. More recently, IEEE 802.1 began work on 802.1CF, whose network reference model work includes

defining interfaces with SDN. Work is also ongoing on enhancements to path control. All of these are components that will be important to SDN and virtualization solutions needed by industry. Often, the challenge in this work is to ensure that other standards bodies are aware of IEEE's work , and that they use this well rather than duplicate the effort, often with less understanding of the technologies IEEE developed. The interaction between 802.1CF and other SDN standards has been a topic of discussion between the OmniRAN Task Group and at least the ONF and the IETF.

In addition to the work on 802.1, many other activities are underway or contemplated within IEEE on SDN. For example, there is a Research Group on Software Defined and Virtualized Wireless access, chaired by Fabrizio Granelli. There is also a Study Goup on Service Virualization, chaired by Mehmet Ulema, that builds on the work of P1903.

Industry Fora for Collaboration

A number of industry bodies enable vendors and users to collaborate on utilizing network technologies. These include bodies such as the BroadBand Forum (BBF), the Metro Ethernet Forum (MEF), and the Optical Interface Forum (OIF). These bodies are effective locales to develop architectures aimed at real-world needs. Many are working on architectures and approaches to using SDN within their problem spaces. While these bodies usual communicate well with more conventional standards activities, the sheer number of industry and standards bodies can make collaboration difficult.

Conclusion

As I have shown, the landscape of SDN standardization is quite broad. But there is duplicate work where there could be reuse. For example, the ONF could have built on the semantics of ForCES, or ForCES could have used the work of the ONF. No one body can possibly do all the work, which means there is room for effective collaboration, based on mutual respect for work done in many places.

As noted in the discussion of ETSI NFV, the easiest way to collaborate is when responsibilities are clear demarcated. This suggests that whenever a standards body undertakes new work, it should look for overlaps and adjacencies, and whenever possible develop clear relationships with other activities. Having common participants working in multiple groups can avoid problems, and, as noted in the discussion of ONFs OF-Config work, can overcome some of the difficulties of collision.

Note also that while it is often attractive for a body to develop its own standards, or even for a group of people to establish their own standards body, this can lead to friction and challenges in the real world where the different protocols must interact and work together to solve real problems. This also hinders interoperability and frequently leads to significantly deployment difficulties.

As has been observed, the tendency in any given body for standards work to expand into adjacent spaces also causes conflict and needs to be managed carefully.

Finally, the emergence of open-source software, an important factor in the standardization landscape, poses its own challenges. It is important that standards bodies and open-source communities look for collaboration and cooperation. Each needs to explain what value it brings to the activities, and respect the value other parties bring. Standards are not implementations, and implementations are not standards. Efforts are further complicated when participants encourage industry to treat open-source code as de facto standards.

The emergence of open-source software, an important factor in the standardization landscape, poses its own challenges. It is important that standards bodies and open-source communities look for collaboration and cooperation.

Acknowledgment

The author would like to thank Eric Gray for his assistance in understanding some of the IEEE activities mentioned in this article.

References

[1] Open Networking Foundation "Software-Defined Networking (SDN) Definition," available: https://www.opennetworking.org/sdn-resources/sdn-definition.
[2] Open Networking Foundation "ONF Overview," available: https://www.opennetworking.org/about/onf-overview.
[3] R. Enns *et al.*, RFC 6241 "Network Configuration Protocol (NETCONF)," Available: http://www.rfc-editor.org/rfc/rfc6241.txt.
[4] M. Bjorklund, RFC 6020 "YANG — A Data Modeling Language for the Network Configuration Protocol (NETCONF)," available: http://www.rfc-editor.org/rfc/rfc6020.txt.
[5] Internet Engineering Task Force information, available http://www.ietf.org.
[6] J. Halpern and J. Hadi Salim, RFC 5812 "Forwarding and Control Element Separation (ForCES) Forwarding Element Model," available: http://www.rfc-editor.org/rfc/rfc5812.txt.
[7] A. Doria *et al.*, Editors, RFC 5810 "Forwarding and Control Element Separation (ForCES) Protocol Specification," available: http://www.rfc-editor.org/rfc/rfc5810.txt.
[8] E. Haleplidis *et al.*, RFC 6053 "Implementation Report for Forwarding and Control Element Separation (ForCES)," available: http://www.rfc-editor.org/rfc/rfc6053.txt.
[9] IETF Charter for the I2RS Working Group, available: http://datatracker.ietf.org/wg/i2rs/charter/.
[10] IETF Charter for the SFC Working Group, available: http://datatracker.ietf.org/wg/sfc/charter/.
[11] ETSI ISG NFV "Our Role & Activities," available: http://www.etsi.org/technologies-clusters/technologies/nfv.
[12] Open Daylight "Why Opendaylight," available: http://www.opendaylight.org/project/why-opendaylight.

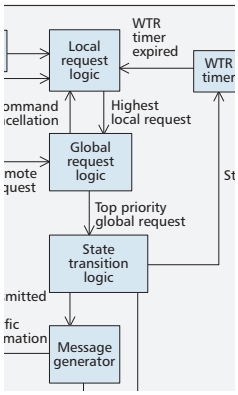
Biography

Joel M. Halpern (joel.halpern@ericsson.com) is a Distinguished Engineer with Ericsson, Inc., with corporate offices in San Jose, California 95134 USA.

MPLS-TP Linear Protection for ITU-T and IETF

The MPLS-TP linear protection specification has resulted in a single unified solution that has been published in IETF RFCs and in ITU-T Recommendation G.8131.

Jeong-dong Ryoo, Taesik Cheung, Daniel King, Adrian Farrel, and Huub van Helvoort



Abstract

The MPLS Transport Profile (MPLS-TP) is a framework for the construction and operation of reliable packet-switched transport networks based on the architectures for MPLS and Pseudowires. Its development has been shared between the IETF, where the MPLS expertise resides, and the ITU-T, with its historic understanding of transport networks. MPLS-TP adds two significant features to the MPLS toolkit: Operations, Administration, and Maintenance (OAM), and linear protection switching. Unlike OAM, which resulted in two application specific and incompatible standards being approved at the World Telecommunication Standardization Assembly (WTSA) in November 2012, the MPLS-TP linear protection specification has resulted in a single unified solution that has been published in IETF RFCs and in ITU-T Recommendation G.8131. This article outlines the novel concepts and operation principles of the unified MPLS-TP linear protection switching mechanism and discusses how it differs from pre-existing solutions. In addition, the issues of compatibility with pre-existing solutions and the applicability to other network topologies are discussed.

COMMUNICATIONS STANDARDS

Introduction

The Multiprotocol Label Switching Transport Profile (MPLS-TP) [1] is designed to be used in an environment that operates with or without an IP-based control plane, meaning that MPLS-TP provides functionality for a centrally controlled transport network (such as in Software Defined Networks (SDN)) or may be integrated with an existing IP/MPLS packet network. In order to enable transport network qualities in an MPLS packet network, MPLS-TP enhances the network's reliability using Operations, Administration, and Maintenance (OAM) to detect and isolate faults, and rapid protection switching (sub-50ms) in the event of failure. These additional functions give the packet network the look and feel of traditional transport networks while building on top of the MPLS architecture.

The roots of MPLS-TP go right back to the original specification of MPLS within the Internet Engineering Task Force (IETF) more than 13 years ago. Since then, MPLS has become an established Internet technology and most packets will traverse an MPLS network somewhere along their end-to-end paths. However, as an Internet technology, MPLS was focused on best-effort routing and connectionless delivery mechanisms. As the possibility arose to utilize the same forwarding hardware in more static environments, it became desirable to add a number of mechanisms to MPLS so that the packet network could perform more like a circuit-switched transport network. The International Telecommunications Union (ITU) used its many years of experience with transport networks to develop requirements and to propose protocol solutions to define what was then called Transport-MPLS (T-MPLS).

As is not uncommon when two worlds collide, the resulting standardization activity resulted in two approaches. Since OAM was the first area worked on, two different, incompatible solutions were developed for MPLS-TP OAM. One is built on pre-existing MPLS diagnostic tools such as Label Switched Path Ping (LSP Ping) and Bidirectional Forwarding Detection (BFD) enhanced through new OAM protocols that can be carried in the MPLS Associated Channel (ACh). The other is developed from the pre-existing ITU-T Ethernet OAM documented in G.8013. Despite very many hours spent debating the merits of the two approaches and the desirability of a single standard for MPLS-TP OAM, agreement could not be reached among ITU-T participants, and the result was the publication of two alternative recommendations: G.8113.1 and G.8113.2.

The next piece of MPLS-TP technology to be worked on was for linear protection switching. As described in [2], linear protection is a protection mechanism that provides rapid protection so that traffic following one path through the network can be switched to a backup path when the working path fails or falls below an acceptable standard, or when an operator command is issued. In a mesh network, linear protection provides a very suitable protection mechanism because it can operate between any pair of points within the network and it can protect against failures in a node, link, transport path segment, or an entire end-to-end transport path. Linear protection relies on a coordination protocol that runs between the end points of the protected path to report errors and to determine what switching actions should be taken. Realizing the problems caused by the existence of two OAM solutions, everyone was particularly concerned to ensure that a single, unified MPLS-TP linear protection protocol and process would be standardized.

After lengthy discussion, Study Group 15 of the ITU-T agreed to develop a single solution for MPLS-TP linear protection that fully meets the ITU-T's requirements by following the normal procedure for creating an RFC in the IETF. Since then, the IETF has made impressive progress toward RFC 7271 [3]. Subsequently, the ITU-T revised G.8131 [4] with the solution specified in [3]. This solution is called Automatic Protection Coordination (APC) in G.8131 [4], and is called "Protection State Coordination (PSC) in Automatic Protection Switching (APS) mode" in [3].

Jeong-dong Ryoo is with ETRI and the Korea University of Science and Technology.

Taesik Cheung is with ETRI.

Daniel King is with Lancaster University.

Adrian Farrel is with Old Dog Consulting.

Huub van Helvoort is with Hai Gaoming BV.

Pre-Existing Solutions for MPLS-TP Linear Protection

Protection State Coordination (PSC)

In bidirectional protection switching schemes, it is necessary to coordinate the protection state between the edges of a protected domain to achieve initiation of recovery actions for both directions: in MPLS-TP this is known as PSC. The requirements for MPLS-TP recovery were worked on jointly by the IETF and ITU-T and are documented in Requirements of an MPLS Transport Profile [5]; they were used to generate the IETF's MPLS-TP linear protection solution [6] known as the PSC protocol.

The purpose of the PSC protocol is to allow communication between an end point at the edge of a protected domain and its peer at the other end of the domain. This communication is used to exchange notifications of the status of the domain and to coordinate the transmission of data traffic. The protocol is a single-phased protocol which implies that each end point notifies its peer of a change in the operation (switching to or from the protection path) and makes the switch without waiting for acknowledgement. Although a single-phase protocol is supposed to complete protection switching via a single message exchange from one end to the other, there are some corner cases for which the exchange of two messages is needed when both nodes have different triggers asking for different paths. Therefore, the protection switching completion time can be delayed up to a message round trip time even in a single-phase protocol.

The developers of the PSC protocol looked to optimize their solution based on the fact that it would only be applied in a packet network, but still looked to re-use many of the concepts familiar in other protection switching systems. However, this led to some significant differences between the protocol messages, state machines, and principle of operation of this approach and those of the APS protocol specified by the ITU-T and used in linear protection for traditional transport networks, such as Synchronous Digital Hierarchy (SDH), Optical Transport Network (OTN), and Ethernet transport networks.

As protection switching operation should complete in one message exchange without any acknowledgement from the other side, the protocol complexity in a single-phased protocol is disposed toward the state machine. Some examples of problematic scenarios of the PSC can be found in [3].

Automatic Protection Switching (APS)

The APS protocol for MPLS-TP as described in [7] is based on the same principles and behavior seen in other ITU-T linear protection technologies. Its implementation has been deployed by several network operators using equipment from multiple vendors. The APS solution was considered in the IETF, but failed to achieve MPLS Working Group consensus. In order to document existing implementations and deployments, this pre-standard solution has been published as an Independent Stream RFC.

The APS for MPLS-TP is consistent with the behavior of Ethernet APS linear protection in G.8031, which has all the necessary functionalities for transport networks and a time-proven approach compared to the PSC. Ethernet APS is also a single-phased protocol, and its state machine had continuously been enhanced until 2011 since its debut in 2006.

Although the APS has all the necessary functionalities for transport networks and a state machine based on the established Ethernet APS state machine, it also reveals inefficient use of network bandwidth to provide protection against Signal Degrade (SD). The existing SD protection mechanism defined in the APS uses a broadcast bridge, which sends traffic to both working and protection paths whenever traffic has to be transmitted to the protection path regardless of the cause of the various protection switching triggers, such as operator commands, signal fail, and signal degrade. This results in inefficient use of network resource and discouraging its use in non-revertive operation.

The round trip time out-of-service issue is unavoidable in a single-phase protocol, but its occurrences are more frequent in the APS than in the PSC. The main reason is that in the basic operation principle of the APS, an end point always sends a No Request (NR) message when a remote message has a higher priority. This basic operation principle is useful to confirm that a request has been accepted by the remote peer, but it also hides any persistent local request. Only after being notified of the clearance of a higher priority remote request, the local node exposes the hidden local request. This might lead to another traffic switching at the remote end.

A Unified MPLS-TP Linear Protection Solution

MPLS-TP Automatic Protection Coordination (APC)

The experience gained during the development of the two solutions described above was used to make a new unified solution for MPLS-TP linear protection. Through the convergence of the PSC and the APS, the APC is now able to solve the aforementioned deficiencies and render improvements on those solutions.

Design Principles of APC

The APC is designed according to the following principles:

Maintain traditional network operational behaviors: For the network operators who have been accustomed to the linear protection schemes seen in other transport networks, bits on the wire and internal mechanisms may not be so meaningful, but maintaining the same operational methods to manage their transport networks is beneficial; it can reduce training costs and simplify operation across multiple transport networks of different technologies.

Define additional mechanisms seen in other transport networks: The additional mechanisms that are essential, but missing from the PSC, have been identified as: an operator command to manually switch traffic from the protection path to the working path (Manual Switch to Working (MS-W)); an operator command to test

The APS solution was considered in the IETF, but failed to achieve MPLS Working Group consensus. In order to document existing implementations and deployments, this pre-standard solution has been published as an Independent Stream RFC.

An important feature of an evolving protocol solution is that it should be backward compatible with deployed equipment, facilitating new equipment to be rolled out incrementally without the need for a flag day across the whole network.

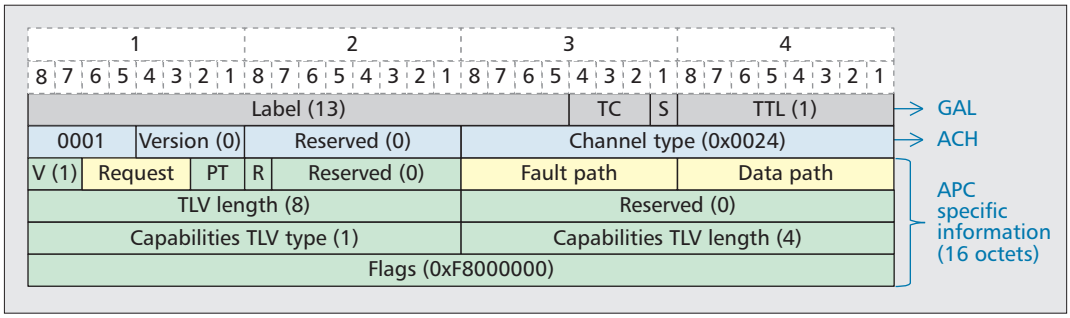


Figure 1. APC PDU format.

protection mechanisms (Exercise (EXER)); and protection switching against an SD defect.

Define an efficient way to provide protection against SD: In the APC, traffic duplication is needed only under SD conditions.

Reuse the basic operation principle of the PSC: The basic operation principle of the PSC, which always reflects its local request in the transmitted PSC protocol messages even when the remote request from the other end has a higher priority, is enforced to reduce the issues of the time for round trip protocol message exchanges in out-of-service cases. Even if the round trip time issue is acceptable in the single phase protocol, it is desirable to avoid.

Reuse the PSC PDU structure: Considering the fact that all the necessary information to perform protection switching is defined for both Protocol Data Units (PDUs) of the APS and the PSC, it is natural to reuse the PSC PDU structure since the PSC achieved IETF MPLS WG consensus and was published as a Standard Track RFC.

Strictly decouple priority evaluation from state machine: In order to define a simple and clean description of the state machine inside each end of a protected domain, priority evaluation for various inputs and state transition table lookup are strictly partitioned.

Reduce possible bugs in state transition tables: By categorizing various inputs carefully and defining a comprehensive operation for each grouped input, any potential bugs in state transition tables can be reduced.

PDU Format of APC

Figure 1 depicts the APC PDU format of G.8131, which is framed in the Generic Associated Channel (G-ACh) as described in the IETF RFC 5586 [8]. Like other MPLS-TP OAM PDUs, the APC specific information is preceded by the four-octet G-ACh Label (GAL) and the four-octet Associated Channel Header (ACH).

Sixteen octets are allocated for the APC specific information. In the APC specific information, all the values except "Request," "Fault Path," and "Data Path" remain the same as configured by the operator. The first two bits are for the Version (V) of the protocol. The Protection Type (PT) field is to indicate the switching type, which can be unidirectional or bidirectional, and the bridge type, which can be a permanent bridge or a selector bridge. The Revertive (R) field is to indicate either revertive or non-revertive operation.

The values of "Request," "Fault Path" and "Data Path" can be changed to provide protection switching against defects and operator commands. They are shown in Table 1.

The remaining fields in the APC specific information are to indicate the protocol capabilities encoded in Capabilities TLV. The description on the Capabilities TLV can be found in [3]. For the protocol operation in G.8131, the fields of the Capabilities TLV should be set as shown in Fig. 1.

APC Process Operation

Figure 2 shows the APC process algorithm, which is performed at both ends of the protected domain. The APC process algorithm is initiated immediately every time one of the input signals changes, i.e. when the status of any local defect (signal fail and signal degrade) changes, when an operator command (lockout of protection, forced switch, manual switch, exercise, etc.) is issued, or when a different APC-specific information is received from the remote end. When there is a need to coordinate timing of protection switches at multiple layers or across cascaded protected domains, a defect may be delayed in the "hold-off timer logic" before being processed.

As multiple local inputs may be active at one time, the "local request logic" determines which of these inputs is of highest priority. The highest priority local input (highest local request) is passed to the "global request logic," that will determine the higher priority request (top priority global request) between the highest local request and the last received remote request.

When a remote APC protocol message arrives, its APC-specific information is subject to the "validity check." By comparing the received APC-specific information with the transmitted, the "validity check" declares a protocol failure if the bridge type or the Capabilities TLV mismatches or the protection switching is not completed within 50 ms. When the remote request specified in the APC-specific information comes to the "global request logic," the top priority global request is determined between this remote request and the highest local request that is present. If the remote request becomes the top priority global request and the highest local request is an operator command, the local command is cancelled.

The top priority global request is then presented to the "state transition logic" to determine the state transition. The consequent actions of the state transition are to set the local bridge and selector positions and to determine the val-

ues of the variable fields for new APC-specific information. If revertive operation is configured, then the Wait-to-Restore (WTR) timer is started to prevent frequent operation of the protection switch due to an intermittent defect. For detailed descriptions of the APC process algorithm, refer to [3] and [4]. Some operation examples of the APC protocol can also be found in [3].

Backward Compatibility

An important feature of an evolving protocol solution is that it should be backward compatible with deployed equipment, facilitating new equipment to be rolled out incrementally without the need for a flag day across the whole network. This section examines how this has been achieved with the unified MPLS-TP linear protection solution, and calls up the evolution of APS as an example of how problems can arise.

APC and PSC

Fundamental to how MPLS-TP linear protection manages to be a unified solution is the way that an implementer can upgrade an existing implementation so that it can support both APC and PSC. Similarly, an operator can introduce APC into a PSC network without breaking anything and continuing to use PSC functionality across the whole network or on the LSPs where one of the end points does not support APC. The main issue to be resolved is that early implementations and deployments of MPLS-TP linear protection are limited to PSC as defined in [6] and need to be able to interoperate with full implementations as defined in [3] and [4].

APC and PSC are defined as operational modes in MPLS-TP linear protection. A mode is a set of capabilities to perform specific functions and to operate in particular ways as indicated in the Flags field of the Capabilities TLV as shown in Fig. 1. For the PSC mode, the Flags value is set to 0×0, and for the APS mode, the Flags value is set to 0×F800000 as shown in Fig. 1. When two implementations use PSC messages to communicate they include the Capabilities TLV that announces the capabilities that they support. Capabilities TLV with other Flags values than 0×F8000000 or 0×0 are treated as an error. MPLS-TP linear protection can only operate if both ends of an LSP announce support for the same mode. Nodes can be configured to support one mode or the other, and this configuration may be per node, per interface, or even per LSP.

A legacy node implemented according to [6] would send no Capabilities TLV since it would be unaware of the new Capabilities TLV: this behavior is taken to mean PSC mode. To facilitate backward compatibility between a legacy and a new end-point, a new node that has the ability to send and process the Capabilities TLV must be able to both send the PSC mode Capabilities TLV and send no Capabilities TLV at all.

APS-2007 and its Upgrade Strategy

Both APC and APS enhance the behavior of the previous version of ITU-T G.8131 (2007), APS-2007, by adding support for the MS-W, SD, and EXER functions, but the APS described in [7] is based on the same protocol and operation principles as the APS-2007. It is common perception that new vision of a protocol would be interoperable with the previous implementation under the limited use of old features. For the network operators who have already deployed the APS-2007, the APS might seem to be beneficial as network upgrade can be done gradually. One good example is ITU-T G.8032 Ethernet Ring Protection [9], where both old and new version nodes can reside on the same ring and provide protection switching with limited functionalities. However, as for the revisions of the APS based linear protection recommendation,[1] the gradual upgrade strategy has not been considered seriously. It is well-recognized among the ITU-T linear protection experts that both end nodes in the protected domain should have the same implementation.

One of the examples that two different versions of the APS protocol lead to a problematic situation is illustrated in Fig. 3. After the signal fail on the working path is recovered, both nodes go to the No Request (NR) state holding traffic on the protection path, which is indicated in NR(1,1) messages. When the NR(1,1) message is received,

Field	Value	Description
Request	14	Lockout of Protection (LO)
	12	Forced Switch (FS)
	10	Signal Fail (SF)
	7	Signal Degrade (SD)
	5	Manual Switch (MS)
	4	Wait-to-Restore (WTR)
	3	Exercise (EXER)
	2	Reverse Request (RR)
	1	Do-not-Revert (DNR)
	0	No Request (NR)
	Others	For future use and ignored upon receipt.
Fault Path	0	Indicates that the protection path is identified to be in a fault condition or is affected by an administrative command, or that no fault or command is in effect on both paths
	1	Indicates that the working path is identified to be in a fault condition or is affected by an administrative command
	2–255	For future extensions and ignored upon receipt
Data Path	0	Indicates that the protection transport entity is not transporting user data traffic (in 1:1 architecture) or transporting redundant user data traffic (in 1:1 under SD on Protection condition or in 1+1 architecture)
	1	Indicates that the protection path is transmitting user traffic replacing the use of the working path
	2–255	For future extensions and ignored upon receipt

Table 1. The values of the fields that vary according to protection switching operation.

[1] ITU-T G.8131 (2007) is consistent with the behaviour of G.8031 (2006) - Ethernet linear protection switching, and the APS mentioned in this article is also consistent with the behaviour defined in G.8031 (2011).

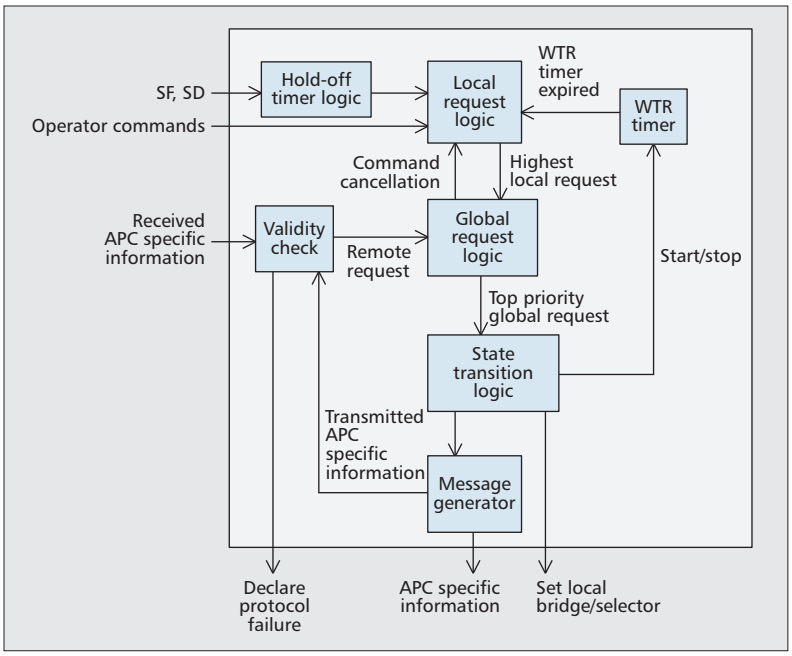


Figure 2. APC process algorithm.

node A running APS-2007 switches traffic to the working path and indicates the No Request state with traffic on the working path in NR(0,0) messages. In the meantime, node Z running the APS starts the WTR timer to see if the clearance of the defect is persistent before reverting traffic to the working path. This results in the traffic discontinuity between two ends during the WTR period which is configured by an operator between five and 12 minutes. As it is obvious that the interoperation between APS-2007 and APS is not possible even for this basic signal fail and recovery scenario, from the perspective of APS-2007 deployment, its upgrade to the APS does not have any real benefits but deviates from the international standard.

Applicability to Other Topologies

As long as what to protect is end-to-end traffic, which is in fact the majority in transport networks, a linear protection mechanism can be used regardless of network topologies. In this section we consider the applicability of the APC to other topologies: ring and shared mesh.

Ring Protection Switching

The applicability of MPLS-TP linear protection mechanisms to ring topologies is described in [10], and the APC can also be used to provide protection of the traffic that traverses an MPLS-TP ring without any new additional mechanisms or protocol. Considering the processing speeds of the current implementations of linear protection processes and OAM sessions, multiple APC processes required to provide protection over a ring would not be a concern. In particular, for the steering architecture, where an ingress ring node determines the forwarding direction between two ring ports, reusing the existing linear protection would be a reasonable choice compared to ring-specific protocols available in other technologies, such as SDH and OTN.

For the other ring protection architecture, wrapping, which has been used in other technologies and which needs to be supported according to [5], the application of linear protection could be quite troublesome and a ring-specific mechanism can be beneficial. It can also be noted that the application of linear protection is limited to a single ring. Interconnected rings cannot be covered efficiently without additional mechanisms, which are not yet available. In the meantime, the benefits of a ring-specific solution also need to be justified against the costs of developing and deploying the ring-optimized solution.

Shared Mesh Protection Switching

In shared mesh protection, the network resources are shared to provide protection for multiple working paths that may not have the same end points. Each working path is protected by a dedicated protection path as in linear protection, but the network resources in a protection path might not be sufficient to simultaneously protect all of the paths for which it offers protection.

One approach can be to define a hop-by-hop restoration mechanism along the protection path. When an end node receives a protection trigger, the end node communicates with each intermediate node along the protection path in a hop-by-hop manner and performs protection switching only after the availability of the shared resources is confirmed by the other end node.

The other approach to achieve shared mesh protection can be to reuse an existing linear protection mechanism for the protection switching action to switch the traffic and define a coordination mechanism to control the use of the shared resources. The additional coordination mechanism focuses on notifying the end points of other working paths not to make any protection switching if the protection resources are insufficient. By separating the traffic switching action from the coordination protocol, which can be done rather more slowly, the protection switching time will be as fast as that in the linear protection. For MPLS-TP networks, the APC protocol can be used as is for the protection switching action for the shared mesh protection.

Conclusions

Linear protection switching is an important component of a circuit-switched transport network enabling the delivery of reliable services even in the event of network faults. Its inclusion in the MPLS-TP portfolio is, therefore, critical to the development of packet switching transport networks based on MPLS technology.

The development of widely agreed and open standards for MPLS-TP linear protection allows all equipment vendors to participate in the market on a level playing field. It also means that network operators can purchase equipment that conforms to a well-known international standard so that it has a high likelihood of interoperating "out of the box." Of course, the fact that there is a single standard, published in their respective ways by both the IETF (the home of MPLS) and the ITU-T (the home of transport networks) makes this situation considerably simpler com-

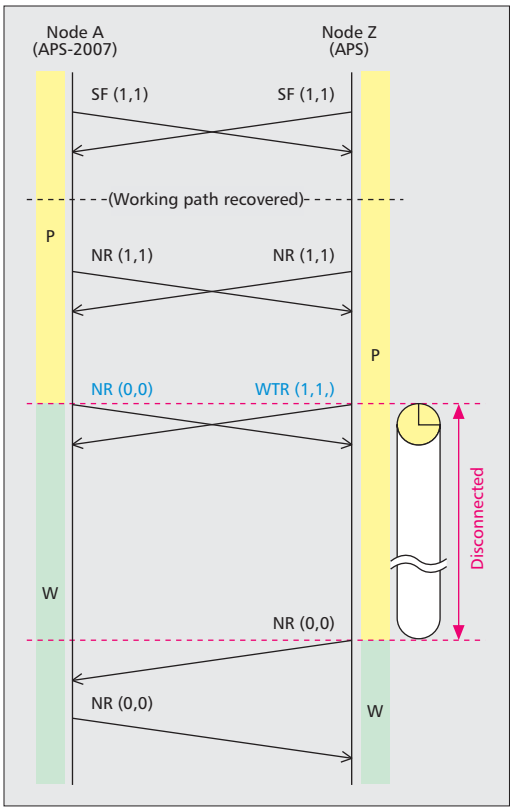


Figure 3. A problem in interoperation between APS-2007 and APS.

pared with, for example, the case of OAM in MPLS-TP networks. Furthermore, the manner in which designers have unified the two desired behaviors within MPLS-TP linear protection to be options within a single protocol specification means that operators are free to choose how to run their networks without having to make detailed technical decisions at the time of purchase.

Finally we observe that the evolutionary approach adopted in the latest MPLS-TP linear protection specifications means that the roll-out of the APC function can proceed incrementally in networks that already deploy PSC, and that operators can make their own deployment choices to support PSC, APC, or both according to their own preferences.

Acknowledgment

This work was supported by the ICT R&D program of MSIP/IITP (Project 10041414, Terabit optical-circuit-packet converged switching system technology development for the next-generation optical transport network).

References

[1] M. Bocci *et al.*, "A Framework for MPLS in Transport Networks," IETF RFC 5921, July 2010.
[2] N. Sprecher and A. Farrel, "MPLS Transport Profile (MPLS-TP) Survivability Framework," IETF RFC 6372, Sept. 2011.
[3] J. Ryoo *et al.*, "MPLS Transport Profile (MPLS-TP) Linear Protection to Match the Operational Expectations of SDH, OTN and Ethernet Transport Network Operators," IETF RFC 7271, June 2014.
[4] ITU-T Rec. G.8131/Y.1382, "Linear Protection Switching for MPLS Transport Profile (MPLS-TP)," 2014.
[5] B. Niven-Jenkins *et al.*, "Requirements of an MPLS Transport Profile," IETF RFC 5654, Sept. 2009.
[6] Y. Weingarten *et al.*, "MPLS Transport Profile (MPLS-TP) Linear Protection," IETF RFC 6378, Oct. 2011.
[7] H. van Helvoort *et al.*, "Pre-Standard Linear Protection Switching in MPLS-TP," IETF RFC 7347, Sept. 2014.
[8] M. Bocci, M. Vigoureux, and S. Bryant, "MPLS Generic Associated Channel," IETF RFC 5586, June 2009.
[9] J. Ryoo *et al.*, "Ethernet Ring Protection for Carrier Ethernet Networks," *IEEE Commun. Mag.*, Sept, 2008, pp. 136–43.
[10] Y. Weingarten *et al.*, "Applicability of MPLS Transport Profile for Ring Topologies," IETF RFC 6974, July 2013.

Biographies

Jeong-Dong Ryoo (ryoo@etri.re.kr) is a Principal Researcher at the Electronics and Telecommunications Research Institute (ETRI) and a UST professor at the Korea University of Science and Technology, South Korea. He holds Master's and Ph.D. degrees in electrical engineering from Polytechnic Institute of New York University, Brooklyn, NY, and a Bachelor's degree in electronic engineering from Kyungpook National University, South Korea. After completing his Ph. D. study in the area of telecommunication networks and optimization, he started working for Bell Labs, Lucent Technologies, New Jersey, in 1999. While he was with Bell Labs, he was mainly involved with performance analysis/evaluation/enhancement study for various wireless and wired network systems. Since he joined ETRI in 2004 his work has been focused on next generation network, carrier class Ethernet and packet transport network technology research, especially participating in OAM and protection standardization activities in ITU-T. He is the co-editor of the G.8131 (MPLS-TP linear protection) and G.8132 (MPLS-TP ring protection) recommendations and a vice-chairman of ITU-T Study Group 15. He co-authored *TCP/IP Essentials: A Lab-Based Approach* (Cambridge University Press, 2004). He is a member of Eta Kappa Nu.

Taesik Cheung (cts@etri.re.kr) is a principal researcher at the Electronics and Telecommunications Research Institute (ETRI), South Korea. He holds B.S., M.S. and Ph. D. degrees in electronics engineering from Yonsei University, South Korea. After completing his Ph. D. study in the area of high-speed circuit design in 2000, he started working for ETRI, where he was engaged in system hardware design such as flow-based QoS switches, Carrier Ethernet switches and packet-optical integrated transport systems. Since 2005 he has participated in ITU-T Q9/15 and contributed to standardization of protection mechanisms for packet transport networks. Since 2010 he has participated in the IETF MPLS WG and contributed to MPLS-TP standardization especially in the area of survivability. His current work focuses on the standardization of MPLS-TP shared mesh protection in the IETF MPLS WG and new protection functionalities such as multipoint Ethernet connection protection and multi-domain segment network protection in ITU-T Q9/15.

Daniel King (d.king@lancaster.ac.uk) is a senior consultant at Old Dog Consulting and is currently studying for his Ph.D. at Lancaster University, where he is researching Network Functions Virtualisation (NFV). He has 16 years of experience working within market leading technology companies. He co-founded Aria Networks with Adrian Farrell and held key roles at Marconi, Movaz Networks, Redback Networks, Cisco Systems, and Bell Labs. Daniel is an active contributor within the IETF, specifically within the PCE, MPLS, L3VPN, and CCAMP working groups, and is an editor or author of numerous IETF Internet-Drafts and RFCs related to path computation, MPLS and network optimization. Daniel is the secretary of two IETF working groups, namely PCE and L3VPN.

Adrian Farrel (adrian@olddog.co.uk) currently serves as one of two routing area directors in the Internet Engineering Task Force (IETF). His responsibilities include the MPLS, CCAMP, L3VPN, and PCE working groups. He is currently funded in this role by Juniper Networks. Adrian has been heavily involved with the IETF for a number of years and is the author of over 50 RFCs. He was among the leaders in the development of some key Internet technologies including GMPLS and PCE. He also runs a successful consultancy company, Old Dog Consulting, providing advice on implementation, deployment, and standardization of Internet Protocol-based solutions, especially in the arena of routing, MPLS, and GMPLS. Adrian is the author or editor of five books on Internet protocols including *The Internet and Its Protocols: A Comparative Approach* (Morgan-Kaufmann, 2004), *GMPLS: Architecture and Applications* (Morgan-Kaufmann, 2005), and *MPLS: Next Steps* (Morgan Kaufmann, 2008).

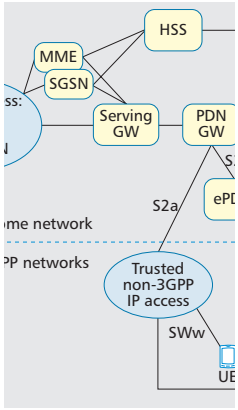
Huub Van Helvoort (huub@van-helvoort.eu) is senior networking consultant for Hai Gaoming BV. He received his MSEE at Eindhoven University of Technology in 1977. He has been a senior member IEEE since 2003. He worked as a designer and architect of SDH, OTN, and PTN equipment at Philips Telecom, AT&T, Lucent, TranSwitch, and since 2005 for Huawei. His is an expert in functional modeling, fault management, performance monitoring, diagnostics and network management. Since 1998 he has been an active participant in transport network standardization: ITU-T, ETSI, ANSI, IETF. He is co-editor of several SDH, OTN, and PTN ITU-T Recommendations for equipment specification, protection, performance monitoring and OAM, and co-editor of MPLS-TP related RFCs. Since 2007 he has been ITU-T Study Group15 rapporteur of expert group Q10: "Interfaces, Interworking, OAM and Equipment specifications for Packet based Transport Networks." He is the author of *The ComSoc Guide to Next Generation Optical Transport: SDH/SONET/OTN* (Wiley/IEEE-Press 2009), *Modeling the Optical Transport Network* (Wiley 2005), and *Next Generation SDH/SONET: Evolution or Revolution?*" (Wiley 2005), and he is co-author of *Optical Transport Networks from TDM to Packet* (ITU-T 2011) and *Optical Networking Standards* (Springer Verlag 2006). See also www.van-helvoort.eu.

The evolutionary approach adopted in the latest MPLS-TP linear protection specifications means that the roll-out of APC function can proceed incrementally in networks that already deploy PSC, and that operators can make their own deployment choices to support PSC, APC, or both according to their own preferences.

Advanced WLAN Integration with the 3GPP Evolved Packet Core

This article explores the advances of the WLAN as an integrated solution using a network-based mobility protocol. This architecture option, also known as S2a by its 3GPP reference point name, is gaining more interest due to its low impact on UEs. Operators can deploy it rapidly, and it offers the possibility of tighter integration between WLAN and 3GPP accesses.

Dinand Roeland and Stefan Rommer



Abstract

The large growth in mobile broadband traffic, wide support of Wireless Local Area Network (WLAN) technology in mobile devices, and the desire of operators to provide services anywhere at any time, have led to a need to integrate WLANs with the 3rd Generation Partnership Project (3GPP) Evolved Packet Core (EPC). This paper starts with a description of an architecture that 3GPP has defined to achieve this. It describes how a mobile device can set up an IP connection over WLAN routed via EPC without imposing new requirements on the mobile device. It also describes how more advanced features that do impose new requirements on the mobile device can be supported in an incremental fashion.

Introduction

Fixed-Mobile Convergence (FMC) is a change in telecommunication systems that has been going on for many years. FMC can have different meanings, but one important aim of FMC is to provide a seamless user experience; i.e., a particular service can be used anywhere at any time. The user need not care where the service is located or by which access technology it is reached. In the last few years one of the main efforts on FMC has focused on integration of WLAN with 3GPP technologies. The vision is a heterogeneous network, where WLAN is integrated with the 3GPP EPC just like any cellular radio-access technology (RAT) [1].

There are a number of key drivers for the integration of WLAN with the 3GPP EPC. First, is the large growth in mobile broadband traffic. To accommodate this, the unlicensed WLAN spectrum can serve as a complement to the cellular licensed spectrum. Second, WLAN access points (APs) are widely available. WLAN can thus be used to complement coverage. Third, there is wide support for WLAN in devices. Most modern mobile devices include both 3GPP radio and WLAN radio. And finally, WLAN integration with EPC offers the possibility of providing end-users with convenient access to mobile broadband services, regardless of the access technology.

Dinand Roeland and Stefan Rommer are with Ericsson, Sweden.

[1] 3GPP uses the terms "trusted" and "untrusted" non-3GPP access when specifying the integrated and the overlay solutions respectively, but these terms are sometimes misinterpreted.

Background to Integration with EPC

3GPP defines a generic architecture to integrate non-3GPP access networks with the 3GPP EPC [2]. A non-3GPP access network uses an access technology different from the 3GPP cellular radio accesses. It can, for example, be WLAN or CDMA2000. In this article we focus on WLAN as the non-3GPP access technology to be integrated with the EPC. Figure 1 shows a simplified picture of such an architecture.

Traffic to and from a mobile device, in 3GPP terminology called User Equipment (UE), can be routed to the EPC via the non-3GPP access using either overlay or integrated solutions.[1]

With an overlay, the UE sets up a secure Internet Protocol (IP) tunnel to the Evolved Packet Data Gateway (ePDG) in the EPC. This is transparent to the non-3GPP access network and thus allows for deployment scenarios where the EPC operator has limited control over non-3GPP access. One example is a home-managed WLAN.

An integrated solution, on the other hand, provides more direct interworking between the non-3GPP access and the EPC and thus allows for scenarios where the EPC operator has more control over the access. An example is an operator-managed WLAN deployment where the WLAN APs are typically not managed by end-users.

COMMUNICATIONS STANDARDS

This article explores the advances of the WLAN as an integrated solution using a network-based mobility protocol. This architecture option, also known as S2a by its 3GPP reference point name, is gaining more interest due to its low impact on UEs. Operators can deploy it rapidly, and it offers the possibility of tighter integration between WLAN and 3GPP accesses.

The next section of this article describes the advances of WLAN integration with the EPC in the 3GPP Release-11 program. Subsequent sections describe further advances made in the 3GPP Release-12 program, and, in particular, the features: handover between 3GPP and WLAN with IP address preservation, attach to non-default APN, EPC offloading and multiple IP connections. The article concludes with an evaluation and discussion of future work, a brief overview of related standardization efforts and a summary.

WLAN Integration with EPC

As part of its Release-11 program, 3GPP launched an effort to define the Trusted WLAN Access Network (TWAN) as an example of a generic integrated non-3GPP access [3]. Figure 2 shows the logical architecture of a TWAN. Note that here the TWAN is a breakdown of the "Trusted Non-3GPP IP Access" of Fig. 1.

Within the TWAN, a number of functions are defined. The WLAN Access Network (WLAN AN) is a collection of WLAN APs. A UE connects to a WLAN AP via the SWw reference point, using the IEEE 802.11 standard. The Trusted WLAN AAA Proxy (TWAP) relays the AAA (Authentication, Authorization and

Accounting) information between the WLAN AN and the 3GPP AAA server in the EPC. Communication between TWAP and 3GPP AAA is performed via the STa reference point. The Trusted WLAN Access Gateway (TWAG) terminates the S2a reference point, which provides a UE access to EPC. S2a uses the network-based mobility protocol General Packet Radio Service (GPRS) Tunneling Protocol (GTP) or Proxy Mobile Internet Protocol (PMIP). Note that the interfaces between the different TWAN functions are not defined in detail. This is not within the scope of 3GPP and may be defined by another Standardization Organization (SDO), e.g., the Broadband Forum (BBF).

3GPP has divided the work on WLAN integration with EPC into two phases. The first phase imposes no new requirements on UEs for WLAN integration. This phase has been completed as part of the 3GPP Release-11 program. By not impacting the UE, this phase allows an operator to achieve WLAN integration with EPC quickly. Figure 3 shows a simplified procedure of a UE attaching to the TWAN.

After selecting a WLAN AP (step 1), the UE performs authentication based on SIM (Subscriber Identity Module) by means of EAP (Extensible Authentication Protocol) using IEEE 802.1X (step 2). Numerous devices already support this, e.g., the devices certified for EAP-AKA prime (Improved EAP Method for 3rd Generation Authentication and Key Agreement) in [4]. The SIM-based method uses the same credentials for accessing 3GPP cellular radio, and provides the user a convenient way to access the network because no manual key or password is needed. Authentication triggers the TWAN to set up the S2a tunnel (steps 3–5). Alternatively, the network may decide to offload this UE from the EPC. Offloading in this context means that user plane traffic via WLAN is not routed via the EPC. Instead, the TWAN routes UE traffic directly to the Internet. An operator may decide to offload the EPC to save user plane resources in the EPC. For offloading, steps 3–5 are omitted. Finally, the UE acquires an IP address (step 7). For EPC-routed traffic, it is the Packet Data Network Gateway (PDN GW) that assigns the IP address/prefix to the UE.

A network-based mobility protocol is used between PDN GW and TWAG so that a point-to-point link is needed between UE and TWAG, as defined in [5]. The assumption for the first phase is not to impose any new requirements on the UE. This has a number of consequences. To avoid new requirements on the UE, it is up to the TWAN to implement the point-to-point link. Furthermore, the UE cannot indicate to the network if the attachment is an initial attach or a handover attach from another access. As a consequence, no handover with IP address preservation is possible in the first phase, despite the fact that a network-based mobility protocol GTP/PMIP is used.

Advanced Features

Lack of support for handover with IP address preservation in the first phase is a functional limitation. 3GPP addresses this and other functional limitations in the second phase of WLAN integration with EPC [3]. The goal is to provide a rich feature set over the WLAN RAT, comparable to existing 3GPP cellular RATs. As this article is written, this phase is being completed as part of the 3GPP Release-12. The next four sections describe the advanced features covered in the second phase.

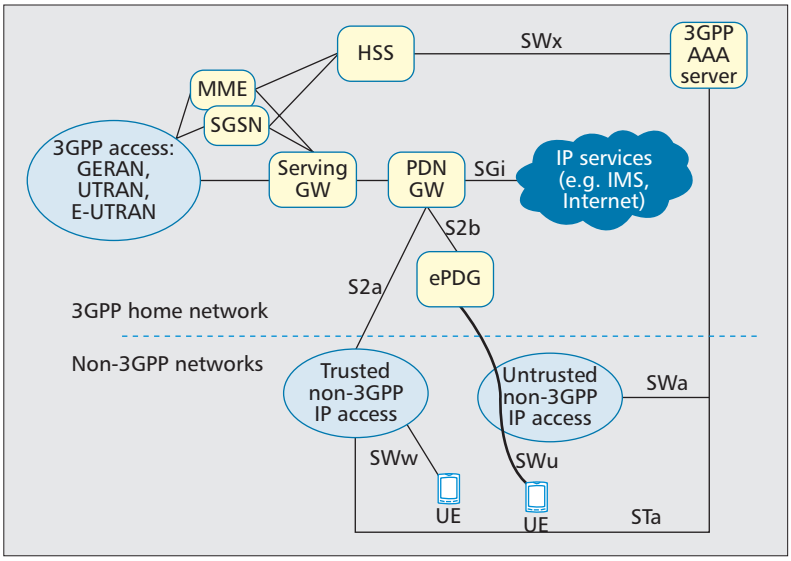


Figure 1. Non-roaming architecture for non-3GPP access connecting to EPC using S2a/S2b.

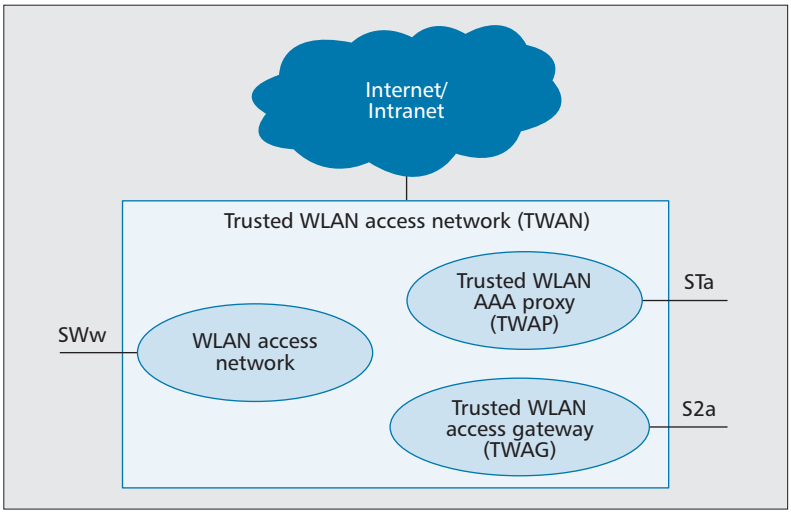


Figure 2. TWAN architecture and functions.

Handover with IP Address Preservation

A typical use case for handover with IP address preservation is an ongoing voice-over-IP (VoIP) call over LTE that is moved to the WLAN. Also other applications, e.g., secure Internet applications, are not designed to handle a change of IP address and would benefit from IP address preservation.

To achieve a handover from a source access to a target access, while preserving the IP address, the UE must be able to indicate that it requests a handover attach instead of an initial

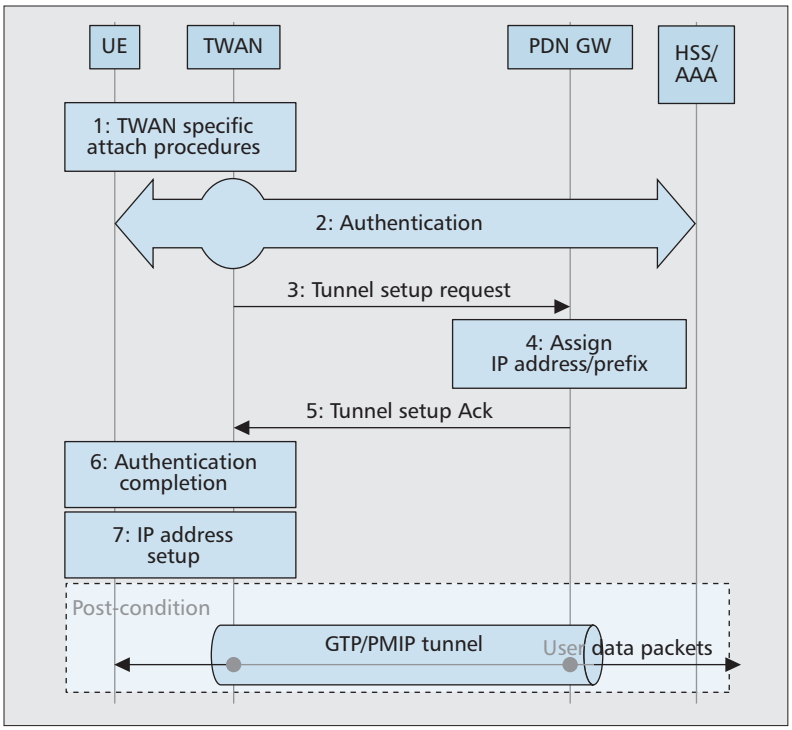


Figure 3. UE attaching to the WLAN network.

attach. One way to achieve this on the WLAN is to indicate handover as part of the authentication. The procedure is shown in Fig. 4.

The difference from Fig. 3 is that a handover indicator is carried in EAP authentication signaling (step 2). This indicator is then also used in the tunnel setup request (step 3), which leads to a handover in the PDN GW (step 4). After this, the PDN GW releases the resources on the source access (steps 8–9).

An important aspect of this procedure is the timing of the different signals. The UE may perform the path switch as part of the IP address setup (step 7). The PDN GW may perform the path switch immediately after setting up the tunnel (step 4). The path switch would then also trigger the tear down of the tunnel in the source access (step 8). Prototype implementations indicate that completing the authentication (step 6) and IP address-assignment procedure (step 7) takes noticeable time. A consequence is that the UE in some cases has not completed the IP address setup (step 7) when the tunnel in the source access is torn down (step 8).

In other words, it is likely that step 8 occurs before step 7. This introduces a time gap in the handover process: the UE has torn down its IP interface on the source access but has not yet set up its IP interface on the target access. If this time gap becomes too large, UE protocol layers may release ongoing sessions. Another consequence is that the PDN GW starts sending IP data towards the WLAN network before the UE has actually completed the IP session setup in the WLAN. This can result in packet loss and extra IP data-interrupt times.

A more optimized handover can be achieved by synchronizing the path switch in the PDN GW with the path switch in the UE. For handover in the other direction, from WLAN to 3GPP RAT, an optimized procedure already exists [6].

Attach to Non-Default APN

3GPP defines the concept of a Packet Data Network (PDN). A PDN is an IP network that a UE attaches to via an access network. In most cases this is the Internet, but it may also be another network like an operator's IP Multimedia Subsystem (IMS) service network or a corporate network. Different PDNs may have overlapping private IP address spaces. The PDN is identified by a character string called an Access Point Name (APN). A PDN GW is a 3GPP operator's gateway to one or more PDNs. A PDN connection is a logical connection between a UE and a PDN GW and is used to gain access to a particular PDN. In the S2a-based architecture explained above, such a PDN connection takes the form of a concatenation of a point-to-point link between UE and TWAG, and an S2a tunnel between TWAG and PDN GW.

For example, a UE has three simultaneous PDN connections, each to a different APN. The first APN may be the Internet, the second an IMS service domain and the third a corporate network. The operator configures its network so that the first PDN connection to the Internet is set up via a first PDN GW. The other two PDN connections are set up via a second PDN GW.

In the first phase, no new requirements are imposed on the UE. Because an APN is a 3GPP concept, not native to the WLAN, the UE cannot indicate to which APN it requests to attach to over the WLAN. When setting up the S2a tunnel, the TWAN then uses the default APN for the UE it receives from the AAA via authentication signaling. In the second phase, this functional limitation is removed. One way to do this is for the UE to provide the APN to the network at authentication (step 2 in Fig. 3).

EPC Offloading

For UE traffic routed via the EPC the TWAN sets up an S2a tunnel to the PDN GW. However, the operator may decide to offload the PDN GW and route UE traffic directly from the TWAN to the Internet (Fig. 2). In such case, the setup of the S2a tunnel (steps 3–5 in Fig. 3) is omitted. The UE still does a SIM-based authentication (step 2) and the UE still acquires an IP address (step 7). For offloaded traffic, it is the TWAN that assigns the address/prefix to the UE. We denote this as an offload connection, as opposed to a PDN connection that is always routed via the EPC. Note that for an offload connection, IP address preservation upon handover is impossible.

In the first phase, the UE is in general unaware if the operator offloads its traffic or routes its traffic via the EPC. Furthermore, the UE cannot send a request to the network about whether it wants an offload connection or a PDN connection. In the second phase, this functional limitation is removed. A way to do this is to let the UE send the offload request in the authentication signaling (step 2 in Fig. 3). Like-

wise, the network can send the offload decision in the authentication signaling to the UE. This way the UE has explicit knowledge of whether its WLAN traffic is offloaded or not. That knowledge can be used by the UE to select the policy rules to apply. An example use case is when the operator configures the UE to send selected traffic via WLAN only when EPC offloading is performed.

Multiple IP Connections

Over a 3GPP RAT, the UE can set up multiple PDN connections simultaneously. A typical use case for this is a UE having one PDN connection to an IMS service domain while having a second PDN connection to the Internet for web browsing. However, in the first phase the UE can set up only a single IP connection over the WLAN, where the IP connection is defined as being either a PDN connection or an offload connection. In the second phase this functional limitation is removed; it is possible for a UE to have multiple IP connections simultaneously over the WLAN. In both the first and second phases, the UE could have one or more PDN connections over a 3GPP RAT, regardless of the number of simultaneous IP connections it has over WLAN.

The solutions for the advanced features described so far are extensions to the authentication signaling. This limits the impact to the UE and the network, making it possible for an operator to deploy these advanced features quickly. However, more fundamental changes are required to support multiple IP connections.

As explained above, each PDN is an independent IP network. For this reason, each PDN connection is treated as a distinct connection between the UE and PDN GW. For example, no traffic is routed from one PDN connection to another within a 3GPP access. In the first phase only a single IP connection over WLAN is supported. Therefore, as explained above, a per-UE point-to-point link between the UE and the TWAG is sufficient. For the second phase we must go one step further. Instead of a per-UE point-to-point link, a per-IP connection point-to-point link between the UE and the TWAG is required. There must be a user-plane traffic separation mechanism to enable the UE and the TWAG to correlate every individual packet to the correct IP connection.

A per-IP connection point-to-point link can be achieved by a tunnel. Such a tunnel can be implemented based on Virtual Media Access Control (vMAC) addresses on the TWAG. Each vMAC corresponds to an IP connection. The vMAC is used in the ordinary IEEE 802 MAC header of frames exchanged between the UE and the TWAG. Other ways to implement such a tunnel include:

- Generic Routing Encapsulation (GRE) [7] on top of the IEEE 802 MAC layer, where the GRE header contains a key that designates the IP connection and where the payload of the GRE frame is the IP packet;
- Virtual Local Area Networks (VLANs) over WLAN [8], where each VLAN corresponds to an IP connection;

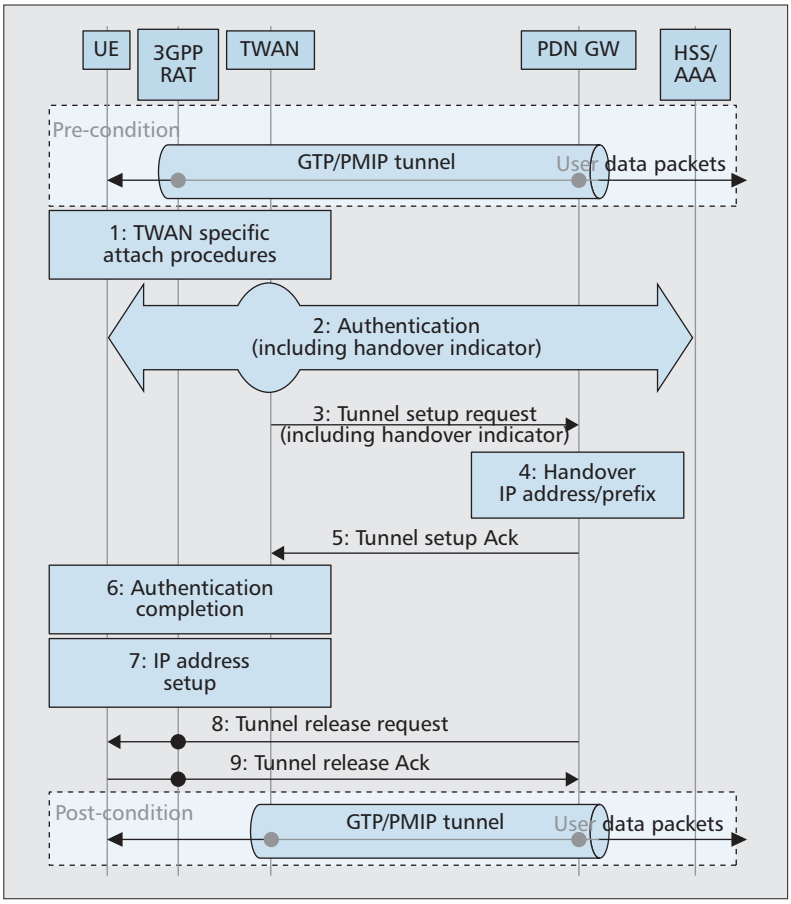


Figure 4. UE performing handover from 3GPP to WLAN.

- Point-to-Point Protocol over Ethernet (PPPoE) [9], where each PPPoE session corresponds to an IP connection.

3GPP has decided to specify the user-plane traffic-separation mechanism based on vMACs. One advantage is that it introduces no extra protocol header.

IP connections need to be managed. For example, it must be possible for a UE to request the setup of an IP connection, and it must be possible for the UE or the network to tear down an IP connection. Because the concept of multiple IP connections, and in particular PDN connections, is not native to the WLAN, a control-plane protocol is needed to manage IP connections. Attachment parameters like handover indicator, APN or offload request can be carried by such a control protocol instead of by authentication signaling.

One way to implement a connection-control protocol is to re-use a subset of the non-access stratum (NAS) protocol [10], the connection control protocol defined for the 3GPP RATs. Other means to implement a connection-control protocol include:

- Extensions to Dynamic Host Control Protocol (DHCP) [11].
- PPPoE.

3GPP has decided to specify a WLAN Control Protocol (WLCP), which is a NAS-like connection-control protocol specifically for WLAN accesses.

The resulting procedure is shown in Fig. 5.

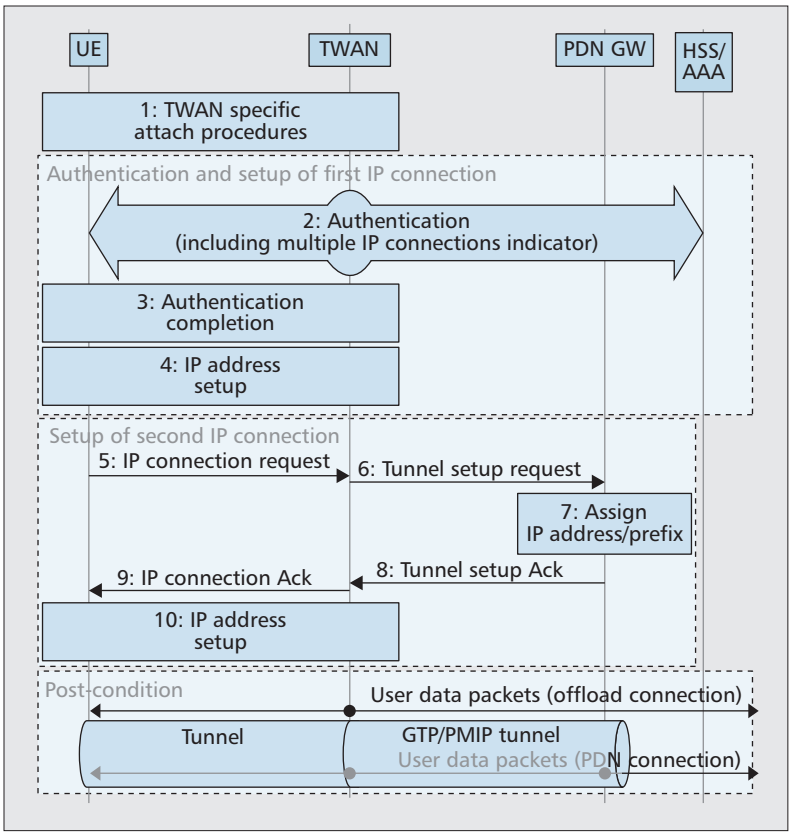


Figure 5. Establishment of an offload connection and a PDN connection.

	S2a phase 1	S2a phase 2 SCM	S2a phase 2 MCM
IP connectivity via EPC	Yes	Yes	Yes
IP session mobility	No	Yes	Yes
Non-default APN	No	Yes	Yes
Explicit offload negotiation	No	Yes	Yes
Multiple IP connections	No	No	Yes
UE impact	No	Limited	Yes

Table 1. Solution evaluation.

As part of the authentication (step 2), the UE indicates a request to set up multiple IP connections. Attach parameters for the setup of the first IP connection may be sent as part of the same step. Because a per-UE point-to-point link can be assumed, no additional per-IP connection tunnel is required for the first IP connection. This allows for a smooth incremental development compared to the first phase and the other advanced features of the second phase. In this example, the UE requests an offload connection as the first IP connection. The TWAN assigns the IP address/prefix to the UE (step 4).

For a second IP connection the UE sends a request using the connection-control protocol (step 5). In this example, the UE requests a PDN connection, which triggers the TWAN to set up an S2a tunnel (steps 6–8). The resulting IP connection is carried in a per-IP connection tunnel between UE and TWAG. This procedure (steps 5–10) is repeated for any additional IP connections.

For multiple IP connections, there is at most one offload connection. The setup of that connection is triggered as part of the authentication signaling. Furthermore, when using multiple IP connections, all PDN connections are set up using the WLCP.

Evaluation and Future Work

The concepts just explained are evaluated in Table 1. The two phases of the S2a-based integrated solution are compared, and for the second phase a distinction is made between Single-Connection Mode (SCM) and Multi-Connection Mode (MCM). The first phase has been standardized as part of 3GPP Rel-11. The second phase is standardized as part of 3GPP Rel-12. All three variants support the basic feature of IP connectivity to the EPC. MCM supports all the advanced features mentioned: handover with IP address preservation, attach to non-default APN, explicit negotiation of EPC offloading, and multiple IP connections. SCM supports all advanced features except for multiple IP connections. Only a single connection over the WLAN is possible. For SCM no WLCP is needed and all the connection setup negotiation is performed by authentication signaling. This limits the impact to the UE.

One topic for future work on WLAN integration with the EPC is support for end-to-end Quality of Service (QoS), in particular when the WLAN access is managed by the operator. A PDN GW can send QoS information to the TWAN via the S2a reference point. On the WLAN transport level, the IEEE 802.11 standard supports the basic mechanisms required for QoS. What is still missing is a feature to perform QoS control signaling between the TWAN and the UE. The WLCP protocol may be extended to support this.

Another topic for future work involves steering a UE's traffic between a 3GPP access and a WLAN access with a granularity finer than APN. For example, 3GPP defines the concept of IP Flow Mobility (IFOM) [2]. This enables some IP flows within a PDN connection to be routed over one access, while other IP flows within the same PDN connection are routed over another access. WLCP may be extended to transport IFOM control-plane signaling.

Related Standardization Efforts

Within the scope of 3GPP, additional efforts are ongoing to achieve a higher degree of WLAN integration with EPC. One topic not discussed in this article is how the UE selects the correct WLAN AP (step 1 in Fig. 3). The WLAN Network Selection for 3GPP Terminals [12] project addresses this. This is related to the ongoing work in the Wi-Fi Alliance with the HotSpot 2.0 specification [4]. Another effort defines the interworking between 3GPP and a fixed broadband access network [13]. This work is performed in cooperation with BBF. As a continuation of this work, 3GPP and BBF are

running a project to come to a further convergence between the EPC and fixed broadband-access networks [14].

Summary

This article describes the background and rationale of integrating the WLAN with the 3GPP EPC. 3GPP has in its first phase defined an architecture for WLAN integration that allows a mobile device to set up a WLAN IP connection that gets routed to the EPC. Attachment to the WLAN network does not impose any new requirements on the UE. This article also explains a second phase of WLAN integration with the EPC, which covers more advanced WLAN features including handover with IP address preservation, attachment to a non-default APN, EPC offloading and support for multiple IP connections over the WLAN. A solution was described to support all these advanced features, except support for multiple IP connections, by a relatively small extension to the authentication signaling. This would allow these features to be implemented rapidly, providing a possibility for an operator to deploy these advanced features rapidly, as well. Finally, a solution was described for also supporting multiple IP connections. This approach can be implemented incrementally to the first solution, giving an operator a smooth upgrade path.

References

[1] Achieving Carrier-Grade Wi-Fi in the 3GPP world, Ericsson Review, 2012.
[2] 3GPP TS 23.402, Architecture Enhancements for non-3GPP accesses.
[3] 3GPP TR 23.852, Study on S2a Mobility based on GTP & WLAN Access to EPC (SaMOG).
[4] Wi-Fi Alliance, http://www.wi-fi.org.
[5] IETF RFC 5213, Proxy Mobile IPv6.
[6] 3GPP TS 23.401, General Packet Radio Service (GPRS) Enhancement for Evolved Universal Terrestrial Radio Access Network (E-UTRAN).
[7] IETF RFC 2890, Key and Sequence Number Extensions to GRE.
[8] IEEE Std 802.11-2012, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.
[9] IETF RFC 2516, A Method for Transmitting PPP Over Ethernet (PPPoE).
[10] 3GPP TS 24.008, Mobile Radio Interface Layer 3 Specification; Core Network Protocols; Stage 3.
[11] IETF RFC 2131, Dynamic Host Configuration Protocol.
[12] 3GPP TR 23.865, WLAN Network Selection for 3GPP Terminals.
[13] H. Kim, L. Kim and A. Kunz, "Enhanced 3GPP System for Interworking with Fixed Broadband Access Network," *IEEE Commun. Mag.*, Mar. 2013.
[14] 3GPP TR 23.896, Technical Report on Support for Fixed Broadband Access Network Convergence.

Biographies

Dinand Roeland (dinand.roeland@ericsson.com) received an M.Sc. cum laude in computer architecture from the University of Groningen, the Netherlands. In 2000 he joined Ericsson, Sweden, as systems manager for core networks. He has worked for Ericsson Research since 2007 and now is senior research engineer. His research interests are in the field of network architectures, with a focus on multi-access networks.

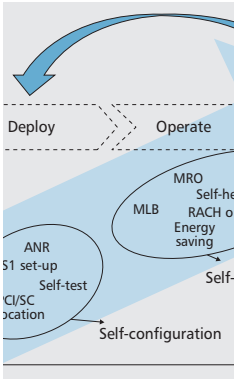
Stefan Rommer (stefan.rommer@ericsson.com) has an M.Sc. in engineering physics and a Ph.D. in theoretical physics, both from Chalmers University of Technology, Sweden. After joining Ericsson in 2001, he worked in different areas of telecom, primarily with mobile-packet, core-network standardization and development. Currently he is a senior specialist in the area of IP mobile networks. Since 2006 he has been active in packet core standardization in the 3GPP.

On the WLAN transport level, the IEEE 802.11 standard supports the basic mechanisms required for QoS. What is still missing is a feature to perform QoS control signaling between the TWAN and the UE. The WLCP protocol may be extended to support this.

Self-Organizing Networks in 3GPP: Standardization and Future Trends

It is important that integrating and operating new and existing network nodes require minimal manual efforts to control OPEX. Consequently, considerable industry momentum has built recently to develop Self-Organizing Network features that can automate mobile network deployment, operation and maintenance.

Ljupco Jorguseski, Adrian Pais, Fredrik Gunnarsson, Angelo Centonza, and Colin Willcock



Abstract

Self-Organizing Networks (SON) is a common term for mobile network automation, critical to the cost-efficient deployment, operation and maintenance of mobile networks. This article provides an overview of SON standardization in 3GPP, including both existing and planned functionalities. It also provides an operator perspective on the relevance and use of 3GPP SON functionalities at different stages of the network design-and-operations cycle. In the long-term it is envisaged that automation will become a natural component in network operations, although the success of SON will depend on automation's benefits in relation to its cost.

Introduction

Recent developments in mobile networks have been driven by the insatiable demand by users for high-speed data. This has led mobile operators to deploy ever more complex networks. In turn, mobile operators now face the challenge of managing these increasingly complex networks comprised of multiple Radio Access Technologies (RATs), different cell types and users with a variety of QoS requirements. At the same time the income of the mobile operators is, typically, decreasing. Thus, it is important that integrating and operating new and existing network nodes require minimal manual efforts to control OPEX. Consequently, considerable industry momentum has built recently to develop Self-Organizing Network (SON) features that can automate mobile network deployment, operation and maintenance.

Within the 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) standardization, SON was among the early system requirements; SON features were included in the first 3GPP LTE release, Release 8 [1, 2]. SON work in 3GPP has been inspired by existing SON studies and the set of requirements defined by the operators' alliance, Next Generation Mobile Networks (NGMN) [3]. In recent literature, SON has been described in [4] and [5]. SON is also related to Minimization of Drive Tests (MDT) [6].

A major driver for mobile operators implementing SON is to decrease CAPEX and OPEX in all phases of the network engineering life cycle: planning, deployment, and operation. The SON features also aim to enhance network performance. The use of SON is crucial, if not inevitable, for most operators running multi-RAT, multi-vendor and multi-layer networks in which an overwhelming number of parameters have to be configured and optimized. For SON to be attractive to mobile operators, its benefits, including both performance improvement and CAPEX/OPEX reduction, should outweigh the cost to implement and manage SON-related functionalities. Towards this goal, operators have a number of high-level objectives for each phase of the network engineering life cycle:

1 **Planning** of new sites (or extension of existing ones) should be as easy, time- and cost-effective as possible, yield the fewest number of sites (or the most cost-effective deployment) for a desired performance, and be based on sufficiently accurate information.
2 **Deployment** of new sites should be as easy as possible with the lowest effort and cost—i.e,"plug and play"—and with no interoperability issues.
3 **Operation** of the network(s) should also be as easy as possible with the lowest effort and cost, allow for quick and effective identification of a problem and its cause, ensure immediate (and preferably automatic) reaction to problems (for example, self-healing and self-optimization), and yield the best possible performance and optimal use of the deployed resources.

An overview of SON functionalities and where they fit in the network engineering life cycle is summarized in Fig. 1. This figure shows that operational efficiency for mobile operators is expected to increase as new SON features become available. By operational efficiency, we mean efficiency in cost and effort spent in planning, deployment and operation, as well as network performance.

This article provides an overview of SON 3GPP standardization, including its relation to MDT and its expected use in the three network engineering phases. First, existing SON solutions in 3GPP (up to Release 11, completed in early 2013) are described. Following that an overview of ongoing SON standardization (i.e., Release 12, expected to be completed at the end of 2014) and a future vision for SON is presented. Finally, we provide a perspective on how these solutions are relevant to mobile operators, including potential challenges.

COMMUNICATIONS STANDARDS

Existing SON mechanisms in 3GPP (up to Release 11)

SON has been discussed as a key enabler of network automation from the very start of work on the LTE Evolved UTRAN (E-UTRAN) specification. The work has materialized as requirements and policies at node behavior level and inter-eNodeB (eNB) signaling procedures, as well as User Equipment (UE) measurements

Ljupco Jorguseski and Adrian Pais are with TNO.

Fredrik Gunnarsson and Angelo Centonza are with Ericsson.

Colin Willcock is with Nokia Networks.

and reports. The E-UTRAN (Fig. 2, right) is maintained and supervised via the network management (NM) system (Fig. 2, left).

The operator interacts with the network at a high level through the NM system, which in turn interacts with the domain manager (DM) through the standardized Interface-North (Itf-N). The DM manages individual network elements (NEs), e.g., eNBs, through the Interface-South, Itf-S.

As described in [2], SON functions can be classified into different types, according to how they are mapped onto the network architecture.

- **NM-centralized SON** operates to meet centralized policies defined in the NM, reconfiguring NE parameters based on network information fed back from the NEs. It is based on the performance indicators and policies defined in the OAM specifications in 3GPP.
- **Distributed SON** is implemented in the NEs (typically eNBs in the case of E-UTRAN, and the Radio Network Controller (RNC) for the UMTS Terrestrial Radio Access Network (UTRAN)). Policies are received from, and KPIs provided to, the NM through the DM over the Itf-N/S interfaces. Inter-NE signaling takes place over standardized interfaces.
- **Hybrid SON** is essentially a combination of both NM-centralized and distributed SON functional components.

In the rest of this section we survey existing SON features and describe them in relation to the management architecture. Specifically, we focus on the additions in Release 11.

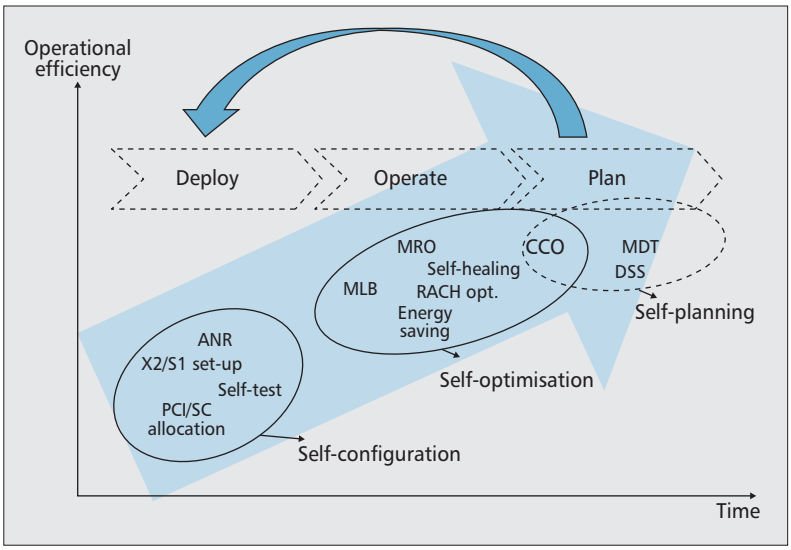


Figure 1. Evolution of SON features.

Self-Configuration

The initial configuration of network elements in a mobile network is complicated by a large number of parameters. Handling configuration manually is tedious and time consuming. This is an obvious candidate for automation because network nodes typically have common values for large portions of the configuration settings..

In self-configuration network elements may be associated with an initial set of site-specific parameters in an optional planning step. This set of parameters may be configured through the 3GPP automatic radio configuration data-handling function (ARCF), and may include cell identities, pre-configured neighbor relations, antenna configurations, transmit power levels, operational carrier, etc. The ARCF, together with any software upgrades, are transferred to the eNB in the self-configuration installation procedure once connectivity has been established. After self-testing, the eNB is operational and ready to serve mobile terminals.

Automatic Neighbor Relations (ANR)

Traditionally, a major configuration/optimization cost for operators has been the manual generation of neighbor relations between cells. This depends on the LTE ANR function located in the eNB. It supports management of neighbor cell relations within E-UTRAN, between E-UTRAN and UTRAN and from E-UTRAN to GERAN and CDMA2000 cells. Based on the UE ANR feature, an eNB or an RNC can request a UE to decode neighbor cell system information and report the decoded information back. Based on this information, the eNB can determine a unique cell identifier for the neighbor cell. This means that the serving eNB has sufficient information to initiate a handover to

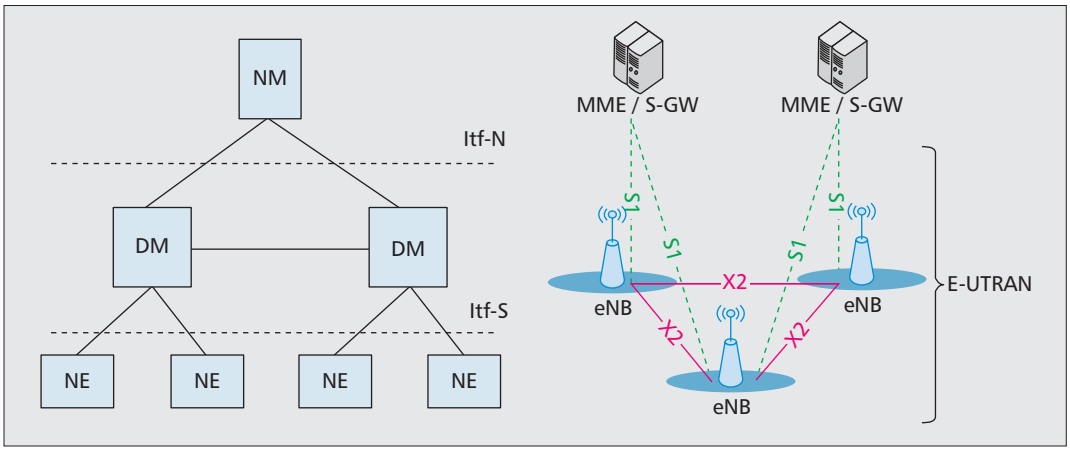


Figure 2. 3GPP network management architecture (left) and E-UTRAN architecture (right).

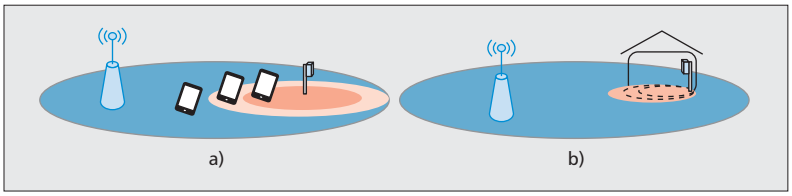


Figure 3. Load balancing between a macro cell and a small cell by adjusting the small cell coverage area via adjustments of a) a range expansion bias, and b) the small cell transmission power.

the discovered cell. Optionally, the eNB may further use the unique cell identifier to retrieve connectivity information from the neighbor base station via S1 eNB/MME configuration-transfer procedures and initiate establishment of an X2 interface. The evident advantage of ANR is that by using UEs to create and update neighbor relations the whole process can be completed automatically. Given the number of UEs in a network, this method is quicker, more reliable and cost effective than drive tests or manual configuration.

Automatic Cell Identity Management

Mobility in 3GPP networks is based on UE-assisted reporting of physical cell identifiers (PCIs) that preferably should be locally unique. Non-unique PCIs can lead to *confusion* (a cell has two or more neighbor cells with the same cell identifier) or *collision* (adjacent cells have the same cell identifier). PCI confusion/collision can be detected via the UE ANR procedure. The OAM system, notified of the detected PCI confusion/collision, can initiate a centralized PCI reassignment mechanism. This proposes a new PCI to the cell based on the neighbor-relation information in the OAM system. Alternatively, the OAM system may provide the eNB with a set of available PCIs to select from, and authorize the eNB to select an alternative PCI, in consideration of assigned PCIs in surrounding eNBs.

Random Access Optimization

The main purpose of the random access procedure is for UEs to notify their presence to the network and establish uplink time synchronization with it. In the procedure, the UE will select a preamble waveform, an access slot and a transmission power. These parameters are subject to optimization to meet requirements in terms of:

- *Access probability*, which is the probability of a UE having completed access after a certain number of random access attempts, or
- *Access delay (AD) probability*, where access delay is defined as the time duration for a random access procedure to complete once it is initiated by a UE.

To assist RACH performance estimation and optimization, the UE can be instructed to provide a RACH report to the eNB after a completed access attempt. This solution is based on UE reports because the UE can monitor radio-related issues which the network may not be aware of. Hence, similarly to the ANR function, this feature makes use of UE monitoring and reporting capabilities.

Mobility Robustness Optimization (MRO)

Robust mobility support is central to mobile networks and MRO is a key SON feature. MRO requirements for intra-LTE mobility are specified in terms of acceptable mobility failure rates while avoiding unnecessary handovers as much as possible. The corresponding requirements for handovers can also be formulated within any RAT or between any two RATs.

The LTE MRO function can be located in the eNB. The handovers are UE-assisted, which means that the UE is configured by its serving eNB to send a Measurement Report (MR) once a reporting criterion is met. Upon receiving a measurement report including information about the candidate cell triggering the report, the serving eNB may initiate the handover procedure to the target cell via X2 or S1 signaling. If the handover fails, the UE will try to re-establish the connection to the radio access network or move to idle mode and reconnect at a later stage.

Recent additions to UE Radio Link Failure (RLF) reports in Release 11 include feedback about the time elapsed since failure (e.g., for removal of stale reports) and information and signaling to detect inter-RAT mobility failures. Similarly, a handover (HO) report can be sent from a different RAT to E-UTRAN to indicate an unnecessary inter-RAT HO. In such cases, upon indication from the source E-UTRAN and after a completed handover, the target RAT configures the UE with inter-RAT measurements of cells in the source RAT (E-UTRAN). If the coverage of one or more E-UTRAN cells is evaluated as acceptable for a specific time after the HO, then the inter-RAT HO is considered unnecessary. The same mechanism allows E-UTRAN to configure a timer in a target RAT to detect inter-RAT ping pongs. Namely, if an inter-RAT HO towards E-UTRAN occurs within such a predefined time window, the HO is considered "too early".

The MRO solution combines events monitored by UEs which are not visible directly from the network together with information from multiple eNBs to detect the root cause of failure. Note that from Release 10, MRO enables UE signalling of RLF Reports after active-idle transitions, which is particularly useful in inter-RAT mobility failure resolution.

Mobility Load Balancing (MLB)

The objective of MLB is to manage uneven traffic distributions, while minimizing the number of needed HOs and redirections. The thresholds triggering an offloading action can be enabled by typical cell overload and related load-performance indicators. To avoid jeopardizing mobility robustness, the same targets specified for MRO can also be considered. The MLB function is in the eNB.

An issue with heterogeneous networks is that small cells may attract too little traffic, which calls for macrocell offloading techniques. One such technique is cell range expansion, where a Range Expansion Bias (REB) is considered for small cells when evaluating measurement-report triggering criteria for some or all UEs (Fig. 3a). The adjustment of the REB can be seen as

mobility load balancing. Such adjustments need to consider UE-specific aspects such as detection capabilities and current traffic-load contributions due to the volatile nature of interference in the range-expansion area. In 3GPP, eNBs can share resource status information via X2. This provides information to select an offload cell and to negotiate mobility parameters via X2 signaling. This design was chosen to enable offloading of UEs, which would otherwise not be retained, to cells with spare capacity without changing radio channel configurations. An alternative offloading technique adjusts the pilot power level of the small cell to increase or decrease its coverage (Fig. 3b).

Energy Savings

Energy is a major cost in operating mobile networks. The only standardized mechanism to reduce energy consumption is to deactivate cells that are temporarily not needed. To facilitate network energy saving, signaling support is specified between base stations as well as between RATs. If an eNB has switched off a particular cell to lower energy consumption, it may notify neighbor eNBs via a deactivation indication over X2. Furthermore, an eNB can request a neighbor eNB to re-activate a previously switched-off cell via a cell-activation request. Release 11 has introduced some inter-RAT support, where it is possible to transfer cell activation/deactivation information between RATs (e.g., UTRAN) via SON transfer messages. The latter approach minimizes complexity while maintaining interoperability.

Minimization of Drive Tests (MDT)

Traditionally, detailed information about actual radio network performance is obtained through drive tests. However, drive tests are costly, time consuming, and typically limited to roads far from where most UEs are located. One attractive alternative is to use UEs as probes that report measurements to the network. In 3GPP, this is referred to as MDT, which was introduced in Release 10, with enhancements in Release 11. MDT is based on the 3GPP trace functionality and enables the operator to configure and initiate trace logging of radio measurements (including RLF Report used for MRO) and optional location information either towards a specific UE or a particular cell or area. MDT is thoroughly described in [6]. Similar to ANR, the rationale behind the MDT design is to be able to use existing UEs in an ad-hoc manner to monitor network behavior and performance. This provides a plethora of statistics to operators without the cost of running dedicated drive tests.

Potential SON Function Conflicts and Resolutions

With many automatic SON functions in the network, there is a concern that they may cause unwanted interactions and even unstable behavior. Potential conflicts between SON functions can be due to how different SON functions affect the same parameter within overlapping time frames. One solution is to rely on standard coordination mechanisms as discussed in 3GPP [8] while another is to ensure that different SON function types are mutually isolated via design principles [9].

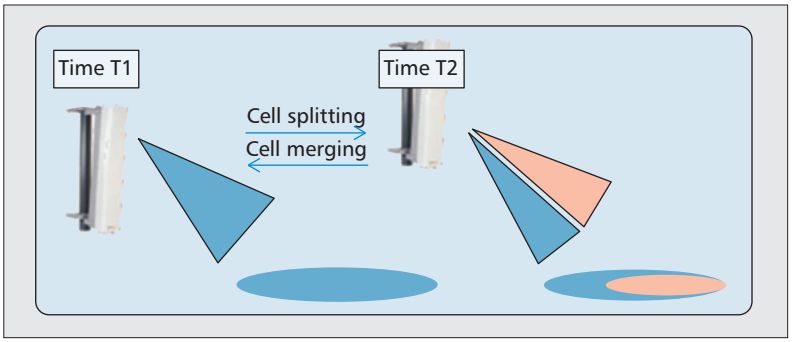


Figure 4. Changing cell pattern of active antenna.

SON for Release 12 and Beyond

Following the work done on SON and MDT in earlier 3GPP Releases, a new study item [7] to extend SON began in early 2013 and concluded in mid-2014. Specification work commenced thereafter and is expected to finish by the end of 2014. In the following sections we detail the Release 12 developments further. Possible developments beyond Release 12 are then also described.

Release 12 SON

Per UE Mobility Differentiation Enhancements — Current specifications enable mobility settings between different UEs to be differentiated. The objective of the "SON for UE types" task is to evaluate if such differentiation can have a negative impact on interoperability. If this is so, then solutions to the interoperability problems are considered. One problem identified is the ping-pong handovers caused by different mobility settings in adjacent cells.

Active Antenna Enhancements — Active antenna systems are one way to increase the capacity of existing networks. Currently, deployments tend to be relatively static, typically just adding vertical sectorization. However, the technology does enable the possibility of more dynamic use, involving UE-specific beam forming, cell shaping, cell splitting and merging. The situation where the number of cells and the cell coverage change over time is shown in Fig. 4. Such merging and splitting can be used to adapt system capacity depending on traffic conditions. It can be seen as a way to provide more flexible coverage/capacity management. However, the ability to merge and split cells dynamically makes the actual management of such systems increasingly complex.

With this in mind the work in Release 12 aimed at enabling support for network deployments based on the generic features of active antennas. More specifically it studied whether existing SON features for deployment automation can be extended to handle dynamic changes possible with active antennas, such as cell splitting or merging. The main focus in the study item concerned connection failures due to cell splitting and merging, as well as impacts on MRO. The work will continue with a Release 13 work item.

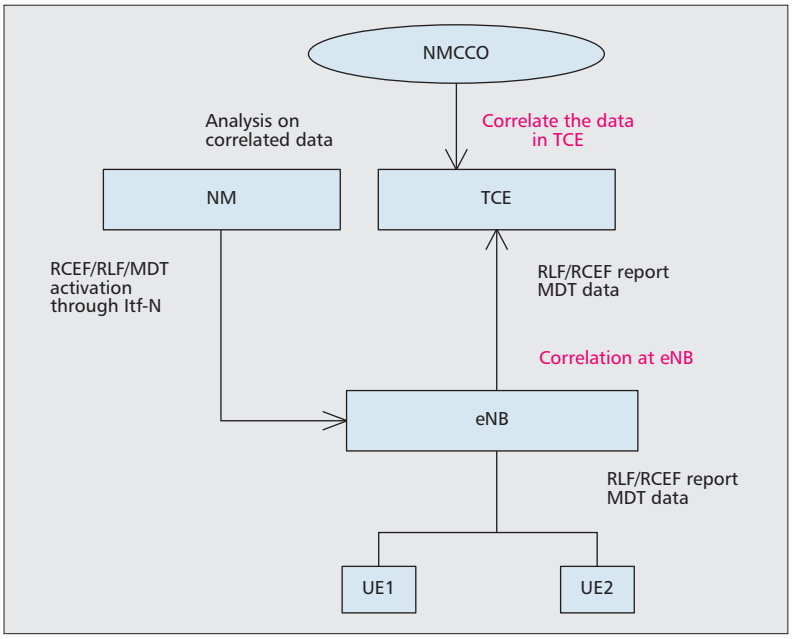


Figure 5. Input data (RLF/RCEF and MDT) for CCO at network management (Fig. 7.3.2.-1 in [10]).

Pre-Release 12 Small Cells Enhancements — The term *small cells* broadly describes operator-controlled, low-power, radio-access nodes. Small cells, which include femto-, pico-, metro- and micro-cells, have a range from tens to hundreds of meters. They're usually deployed by operators wherever additional capacity is needed. Specific SON functions for small cells may reduce network-planning efforts, enhance network optimization and address problems and scenarios specific to small-cell deployment.

Mobility robustness is a challenge, especially because moving UEs may switch rapidly among small cells. The proposed Release 12 enhancements are intended to provide the network additional information (e.g., further RLF reports if failure occurs after re-establishment and UE time-to-trigger (TTT) information), which can be used during MRO analysis so that better corrective actions can be taken. Additionally, S1- and OAM-based solutions have been proposed to simplify inter-RAT RLF reporting in "LTE island coverage" scenarios where there is no LTE coverage surrounding the small cells.

NM-Centralized Coverage and Capacity Optimization and Coordination — The NM-centralized coverage and capacity optimization (CCO) function was facilitated in Release 10 and Release 11 by standardizing activation and reporting of measurement traces reported by the UE or the eNB, including MDT data, radio-link failure (RLF) events, and RRC re-establishment failure (RCEF) events. These traces are then stored at the Trace Collection Entity (TCE) and processed by the CCO functions at the NM level for discovering capacity or coverage issues (Fig. 5). In Release 12, the CCO work focused on anonymous data collection to protect user privacy and correlation of the data from the UEs and eNB either at the eNB level or TCE level (Fig. 5). The targeted use cases were discovering coverage holes and capacity issues in LTE and UMTS, adapting the cell coverage to the user spatial traffic demand, discovering LTE coverage holes via underlying UMTS coverage, etc., as listed in [10].

Multi-Vendor Network Element Plug and Play — The Release 12 work item covered scenarios where an eNB is connected to the secure operator network either via an external network or a non-secure operator network. Server addresses needed for various configurations are obtained via domain name servers [11].

Future Directions

The current work on Release 12 is likely to introduce a number of new features that may well require refinements and additions to existing SON and MDT functionality. One example is Release 12's small-cell enhancements, such as Dual Connectivity [12], that could lead to significant changes to the overall 3GPP Radio Access Network architecture and operation. Another example is 3GPP-WiFi integration. These envision mass deployment of small cells to increase system capacity and user throughput, which by definition requires automated deployment and management for costs to be acceptable to the operators.

The practical and market issues from deployed networks also shape the future directions for SON and MDT development. Changes in user behavior and expectations can cause new problems or requirements so that operators might aim at optimizing quality of experience (QoE) for particular services and users in specific conditions. New deployment trends (e.g., network sharing, small cells, dual connectivity, multicast and broadcast data transfer, or device-to-device communication) can highlight issues or increase the priority for SON and MDT solutions in specific areas.

Research work also leads the way to new ideas and functionality for SON and MDT. The European FP7 SEMAFOUR project[1] [13] is one relevant research project that aims at a unified self-management system. This would enable network operators to holistically manage and operate their complex heterogeneous mobile networks. The ultimate goal is a system that enables an enhanced quality of user experience, improved network performance, improved manageability and reduced operational costs. An envisioned future of the SON system is shown in Fig. 6.

The key new element is the integrated SON management layer. This includes a policy-translation layer which converts high-level goals down to individual SON functions. Further, there is a centralized or decentralized SON-coordination functionality to avoid conflicts between individual SON functions, and a set of powerful new multi-RAT, multi-layer SON functions to handle the complex mobile-communication networks of tomorrow. A decision support system (DSS) [14] is envisioned for automatic generation of recommendations for relevant network extensions based on current and desired performance KPIs, available options for network extensions (e.g., reconfiguration or extension of existing sites,

[1] The research leading to these results has received funding from the European Union, Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 316384.

adding new sites, etc.), projections of future traffic trends and cost constraints.

Operator Perspective on SON

The trend in network operations is to gradually move from "semi-manual" toward autonomous planning, deployment, and optimization (Fig. 1). A semi-manual, or open loop, operation means that SON functionalities suggest configurations which are first approved by the operator before being implemented. Autonomous network operation, also known as closed-loop, means that approval by the operator is skipped. Instead, the operator simply defines high-level performance goals (in the form of a policy) and monitors to what degree the policy is satisfied in the network.

In the **planning phase**, the CCO and DSS functions, with the support of MDT, can reduce the operator's effort in planning (i.e., reduce OPEX) and selecting optimal network extensions (i.e., reduce CAPEX). Operators will still need an initial planning effort to deploy the coverage layer, but this effort will diminish as the coverage layer is enlarged (or completed) and has to be extended with a capacity layer. It is expected that the CCO and DSS functions will be centralized and will typically run at the NM level. This is because CCO and DSS analyse and optimize a cluster of base stations and the dynamics of the recommended reconfigurations or extensions are relatively slow, e.g., up to a few reconfigurations per day or week for CCO or long-term extensions deployed over several months for DSS.

To include NM-centralized SON in the planning phase successfully, operators must address the following challenges:

1 **Availability and accuracy of input data for proper NM-centralized SON decision-making.** Input data provided to NM-centralized SON functions may be in the form of MDT and KPI reports/traces, enriched with geographic coordinates. Collecting this data is facilitated by UE and eNB features that are only optional. Due to this and the existence of legacy UEs and base stations, the availability of input data may be limited. A considerable portion of input data may be provided by vendor-proprietary solutions.

2 **Facilitating the collection and processing of data.** Due to the large number of base stations and users, as well as frequent logging and reporting of relevant information, a huge amount of data needs to be handled by the operator's network management system. This requires sufficient Itf-N transport network capacity, data storage capacity and processing capacity for the NM-centralized SON algorithms.

3 **Linking and synchronizing the NM-centralized SON functions functions with an operator's existing planning tools and processes, as well as BSS/OSS systems.** It is important that the same, up-to-date input data is available to all tools/processes involved in the planning phase. Discrepancies in input data among different processes and tools during planning might result in sub-optimal configurations and network instabilities.

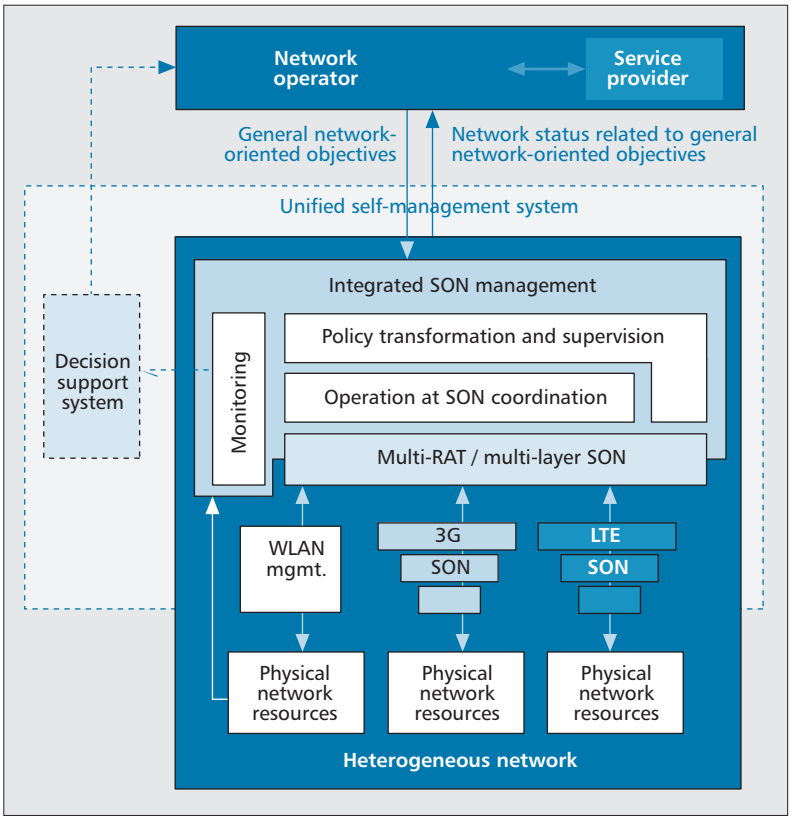


Figure 6. SEMAFOUR vision for future SON system.

In the **deployment phase**, the self-configuration functions enable operators to install new nodes (including Home eNBs) in plug-and-play fashion. Operators' effort in configuring and optimizing intra- and inter-LTE neighbors is reduced (or ideally completely avoided) by the utilization of ANR. Because the LTE neighbor relation establishment includes automatic X2 setup this phase has to be tested in the case of a multi-vendor deployment with neighboring base stations from different vendors. Additionally, the effort in PCI allocation is also avoided because assigning PCIs to cells is automated with the assistance of ANR, and coordinated via the central OAM system, as explained in Section 2. Note here that in case of multi-vendor and small cells deployment, the PCI assignment among the different vendors' nodes and layers has to be coordinated. The self-configuration development of Release 12 can facilitate multi-vendor eNB plug-and-play support that is especially important when deploying network elements with backhaul outside the secure network of the operator.

In the **operations phase**, the distributed SON functions (e.g., MRO, MLB, Energy Saving, etc.) enable operators to have cell-specific and dynamic configurations (e.g., typically few changes per hour), in contrast to base station cluster-based, slowly varying configurations in the planning phase. Cell-specific and dynamic configurations are even more applicable for small cells deployments because (optimal) configuration of small cells and surrounding macro-

SON functions standardized up to 3GPP Release 11 include self-configuration and self-optimization features. These features are supported by MDT, which facilitates the gathering of measurements from UEs. In 3GPP Release 12, to be completed in December 2014, further SON enhancements have been investigated.

cells depends on local traffic and radio- propagation conditions. Because these SON functions operate on a short time-scale and are based on local conditions, they are typically deployed in a distributed or hybrid architecture. Consequently, for multi-vendor deployments in which X2 signaling messages are exchanged, agreements on parameter configurations between neighbor eNBs are crucial for the self-optimization SON functions [15] to work properly. One example is the exchange of load-level information and handover-cause information between eNBs, which has been subject recently to alignment agreements in Release 12.

Conclusions

Network automation in general and SON specifically provide the most promising paths for mobile network operators to handle the increasing pressure to provide ever-higher-performing services, while reducing costs at the same time.

A considerable number of SON functions have been standardized in 3GPP to facilitate automation of planning, deployment and optimization for mobile operators. There is a clear trend that network operations are shifting from manually intensive and static (or slowly changing) network configurations toward (semi-) automated and dynamic/pro-active network operations.

SON functions standardized up to Release 11 include self-configuration and self-optimization features. These features are supported by MDT, which facilitates the gathering of measurements from UEs. In Release 12, to be completed in December 2014, further SON enhancements have been investigated. These can enable operators to have tailor-made optimization based on UE groups, facilitate dynamic shaping of the cell coverage area (including cell splitting) for eNBs equipped with adaptive antenna systems, and assist in the planning phase with the deployment of a NM-centralized CCO function.

Operators must meet several challenges if they're to incorporate NM-centralized SON functionalities successfully. First, the operator's management system should be able to handle the huge amount of data needed for the different SON functions and ensure that the data is synchronized and up-to-date in all supporting tools and processes. Second, although there has been a considerable amount of SON standardization, vendor-proprietary solutions due to legacy UEs or network systems (e.g., GSM and UMTS) will be needed to facilitate the SON functions, which might add to complexity in multi-vendor deployments.

References

[1] Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (EUTRAN); Overall Description; Stage 2, 3GPP Standard TS 36.300, v. 12.1.0.

[2] Telecommunication Management; Self-Organizing Networks (SON); Concepts and Requirements, 3GPP Standard TS 32.500, v. 11.1.0.

[3] NGMN, "Recommendation on SON and O&M Requirements," 2008.

[4] O.G. Aliu *et al.*, "A Survey of Self Organization in Future Cellular Networks," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, 2013, pp. 336–61.

[5] M. Peng *et al.*, "Self-Configuration and Self-Optimization in LTE-Advanced Heterogeneous Networks," *IEEE Commun. Mag.*, vol. 51, no. 5, May 2013, pp. 36–45.

[6] J. Johansson *et al.*, "Minimization of Drive Tests in 3GPP Release 11," *IEEE Commun. Mag.*, vol. 50, no. 11, Nov. 2012, pp. 36-43.

[7] Study on Next Generation SON for UTRA and LTE, 3GPP Technical Report TR 37.822, v 1.4.0.

[8] Telecommunication Management; Self-Organizing Networks (SON); Policy Resource Network Model (NRM). 3GPP Technical Specification TS 36.522, v11.7.0.

[9] K. Zetterberg *et al.*, "On Design Principles for Self-Organizing Network Functions," IWSON at ISWCS 2014, Aug. 2014.

[10] Telecommunication management; Study on Network Management (NM) Centralized Coverage and Capacity Optimization (CCO) Self-Organizing Networks (SON) function. 3GPP Technical Report TR 32.836, v12.0.0.

[11] Telecommunication Management; Procedure Flows for Multi-Vendor Plug-and-Play eNode B Connection to the Network (Release 12), 3GPP Technical Specification TS 32.508, v12.0.0.

[12] Scenarios and Requirements for Small Cell Enhancements for E-UTRA and E-UTRAN (Release 12), 3GPP Technical Report TR 36.932, v12.1.0.

[13] SEMAFOUR: Self-Management for Unified Heterogeneous Radio Access Networks, http://www.fp7-semafour.eu/, accessed 2 May 2014.

[14] A. Eisenblatter *et al.*, "Integrated Self-Management for Future Radio Access Networks: Vision and Key Challenges," Future Network and Mobile Summit 2013, July 2013.

[15] 4G Americas, "Self-Optimizing Networks in 3GPP Release 11: The Benefits of SON in LTE," Oct. 2013.

Biographies

Ljupco Jorguseski received a Dipl. Ing. degree in electrical engineering from Ss. Cyril and Methodius University, Skopje, Republic of Macedonia, in 1996 and a Ph.D. degree in 2008 from Aalborg University, Denmark. From 2003 he has been a senior consultant, wireless access at TNO (Netherlands Organization for Applied Scientific Research), in Delft, focusing on radio planning and self-optimization of wireless networks, including 3GPP standardization. He has co-authored more than 15 scientific papers and book chapters, and has patents pending.

Adrian Pais is project manager and consultant at TNO where he contributes to research and consultancy projects in wireless networks. He has BE with honors and Ph.D. degrees from The University of Auckland, New Zealand. His most recent technical interests include device-to-device communication, self-organizing networks, network energy savings, and 5G networks. Adrian represents Dutch mobile operator KPN in 3GPP RAN standardization. He currently serves as a director on the board of the IEEE Foundation, the philanthropic arm of the IEEE.

Fredrik Gunnarsson [SM'14] received the M.Sc. and Ph.D. degrees in electrical engineering from Linköping University, Sweden, in 1996 and 2000, respectively. In 2001, he joined Ericsson Research, Linköping, where he is currently a senior specialist in radio self-organizing networks (SONs). He is also an associate professor in automatic control at Linköping. His research interests include signal processing and automatic control aspects of RRM and mobile localization. He is associate editor of *IEEE Transactions on Vehicular Technology*.

Angelo Centonza obtained his B.Sc. and M.Sc. degrees with honors in electrical engineering in 2002 from Politecnico di Bari, Italy. He went on to obtain a Ph.D. in hybrid broadcast/telecommunication networks at Brunel University, London, UK. After working in the areas of IEEE/3GPP standardization and telecommunication systems for defense applications, Angelo joined Ericsson Research in 2011 where he is involved in the research of SON, HetNet and small cells. He is also a 3GPP standardization delegate.

Colin Willcock is head of radio network standardization at Nokia Networks, Munich. He received a B.Sc. degree from Sheffield University, UK, in 1986, an M.Sc. from Edinburgh University in 1987, and a Ph.D. in parallel computation from the University of Kent, also in the UK, in 1992. Colin was part of the core ETSI team that developed the TTCN-3 language and spent many years leading TTCN-3 language development. In the past, he has worked on numerous standardization efforts at ETSI, ITU-T, and 3GPP. He also has also led other European research projects. Currently he is leading the FP7 SEMAFOUR project, which is developing the next generation of SON solutions.

Medical Devices

The IEEE Standards Association (IEEE-SA) develops standards that enable the "on the go" technology of mobile devices used by clinicians and caregivers to help patients live active and independent lives.

The IEEE Engineering in Medicine and Biology (EMB) Committee and the Personal Health Devices (PHD) Working Group develop standards that help enable personal health devices to plug and play with mobile phones and home hubs, using Bluetooth and USB specifications. These standards support mHealth initiatives that couple devices, communication, and health/wellness/medical communities.

The IEEE-SA PHD Working Group, together with the IEEE-SA Upper Layer (UL) and Lower Layer (LL) Working Groups, comprises the IEEE 11073™ family of standards. These standards are coordinated across the entire healthcare continuum for personal health device communications for monitors, point-of-care communication foundations, and transport profiles.

The IEEE 11073 standards assist in the support of patients living independently with chronic diseases like asthma, diabetes, congestive heart failure, chronic obstructive pulmonary diseases, high blood pressure, stroke and atrial fibrillation.

Working together for the greater good

IEEE is dedicated to advancing technological innovation and excellence as is your gateway to the most vital information in the healthcare industry today, connecting you with a global network of experts in biomedical engineering, consumer electronics, personal health devices, and more.

IEEE LifeSciences: a technical community of students, professors, researchers, engineers, scientists, medical practitioners, and professionals to advance the application of engineering and technology to the life sciences. www.lifesciences.iee.org

IEEE Committee on RFID: revolutionizing healthcare delivery and optimizing workflow. www.ieee-rfid.org

IEEE Communications Society: publications, conferences, educational programs, local activities, and technical committees that foster original work in all aspects of communications science, engineering, and technology. www.comsoc.org

IEEE Computer Society: technology information, inspiration and collaboration. www.computer.org

IEEE Engineering in Medicine and Biology Society: a driving force in biological, medical, and healthcare brining together the people, practices, information, ideas and opinions that are shaping one of the fastest growing fields in science. www.embs.org

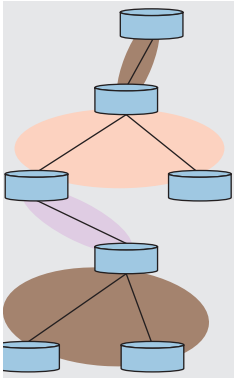
Get more information on Healthcare IT standards

standards.ieee.org/findstds/standard/healthcare_it.html

6TiSCH: Deterministic IP-Enabled Industrial Internet (of Things)

In November 2013 the new 6TiSCH IETF working group was created to "glue" together a link-layer standard offering industrial performance in terms of reliability and power consumption, and an IP-enabled upper stack. This working group is standardizing mechanisms for 6TiSCH networks to operate at the trade-off between throughput, latency, and power consumption most appropriate for the application, while maintaining ultra-high reliability.

Diego Dujovne, Thomas Watteyne, Xavier Vilajosana, and Pascal Thubert



COMMUNICATIONS STANDARDS

Abstract

Industrial and IP-enabled low-power wireless networking technologies are converging, resulting in the Industrial Internet of Things. On the one hand, low-power wireless solutions are available today that answer the strict reliability and power consumption requirements of industrial applications. These solutions are based on Time-Synchronized Channel Hopping, a medium access control technique at the heart of industrial standards such as the WirelessHART and ISA100.11a, and layer 1 and 2 standards such as IEEE802.15.4e. On the other hand, a range of standards have been published to allow low-power wireless devices to communicate using the Internet Protocol (IP), thereby becoming true "fingers of the Internet," and greatly simplifying their integration into existing networks.

This article acknowledges the standardization effort to combine those capabilities. The networks resulting from this convergence exhibit reliability and power consumption performances compatible with demanding industrial applications, while being easy to integrate, and following the end-to-end paradigm of today's Internet. In particular, this article presents the work being done in 6TiSCH, a newly-formed working group in the Internet Engineering Task Force, which is standardizing the mechanisms making the Industrial Internet of Things a reality.

Introduction

Steel mills, oil refineries, and offshore drilling platforms implement complex industrial processes, which require a tight control and a scalable diagnostic transport. Thousands of sensing points are used to report temperature, pressure, and tank fill levels to an industrial process control center. This center, either in an automated way or through human intervention, uses that information to control an actuator, start a new production stage, schedule maintenance, or trigger an alarm. Communication between sensors, actuators, and the control center is done through an industrial network. This class of network needs to offer ultra-high reliability, while operating reliably in harsh environments. Network failures can have catastrophic consequences, and are therefore not an option. In order to gain higher security and reliability, an industrial network is classically partitioned in a hierarchical manner per the Purdue Enterprise Reference Architecture (PERA), using different technologies at each level, and wireless networks are used at the last hop(s) to the field devices. Industrial networking technology has developed over the last 40 years to satisfy those requirements. A major industrial standard is HART [1], a set of standards covering the protocol, connectors, and wires interconnecting the different networked elements.

Depending on the safety regulations in use, the price of drawing cables across an industrial plant can run from $100s/ft to $1,000s/ft. Planning, installing, and maintaining these cables represents a large portion of the cost of ownership of such a wired industrial network. As detailed in this article, advances in reliable wireless technology enable low-power wireless networks to exhibit 99.999% end-to-end reliability and a decade of battery lifetime [2], making them a suitable alternative to wires. This has triggered a trend for industrial networks to "go wireless."

Since 2007, a constant standardization effort, in particular at the IETF, has enabled constrained wireless devices to behave as regular Internet hosts, by acquiring IPv6 addresses, forming multi-hop meshes, and interacting with Internet clients and servers through standard application-layer protocols. Such networks are much easier to integrate in a production system, most of which have already moved toward an IP-based architecture. This ease of integration has triggered a trend for industrial networks to "go IP."

The development of wireless, IP-enabled industrial networks is a factor in the convergence between industrial networks and traditional networks, a trend known as "IT/OT convergence." Operational Technology (OT) refers to industrial networks, which focus on highly reliable, secure, and deterministic networking. Information Technology (IT) refers to the Internet, which relies on selective queuing and discarding of packets to achieve end-to-end flow control. The goal of the IT/OT convergence is to leverage IT technologies to solve OT problems, for instance, by applying the concept of Device Virtualization to emulate field logic controllers and instrumentation and simplify the deployments, or use Big Data/Analytic techniques, operating on large historical repositories as well as vast amounts of live feeds from large scale monitoring installations, to optimize industrial processes, effectively implementing the concept of the Industrial Internet.

An indication of this trend is the creation in 2014 of the Industrial Internet Consortium,[1] a non-profit partnership of industry, government, and academia, created to accelerate the development and availability of intelligent industrial automation.

In November 2013 the new *6TiSCH* IETF working group was created to "glue" together a link-layer standard offering industrial performance in terms of reliability and power consumption, and an IP-enabled upper stack. This

Diego Dujovne is with Universidad Diego Portales.

Thomas Watteyne is with Linear Technology, Dust Networks Product Group.

Xavier Vilajosana is with Worldsensing.

Pascal Thubert is with Cisco Systems.

[1] http://www.iiconsortium.org/

working group is standardizing mechanisms for 6TiSCH networks to operate at the trade-off between throughput, latency, and power consumption most appropriate for the application, while maintaining ultra-high reliability. 6TiSCH is expected to become the standard for low-power wireless industrial monitoring applications.

This article provides an overview of the trends of industrial networking standardization activities toward wireless and IP technologies, and summarizes the ongoing standardization activity of the IETF 6TiSCH working group. We highlight the challenges faced by low-power wireless networks. We detail the standards developed to answer those challenges, and the resulting commercial products dedicated to industrial wireless. We present the goals and standardization activities in the IETF 6TiSCH working group, including different considered scheduling approaches. Finally, we conclude this article.

The Promise of Wireless

Wireless technology enables "peel-and-stick" deployment and requires no maintenance of cables, drastically reducing associated costs. In addition to removing the need for cables for communication, low-power wireless devices are typically powered by a combination of batteries and energy harvesting solutions, thereby also removing the need for power cables. The lifetime of the device is related to its average power consumption, so ultra low-power devices are needed.

Transmit power, modulation, and data rate all influence the power consumption of a low-power wireless device. The IEEE802.15.4 [3] standard was first published in 2003, and offers a healthy trade-off between transmit power (0–10dBm is typical), data rate (250kb/s at 2.4GHz), and maximum packet size (127 bytes). Although the standard was revised twice, the physical layer it defines hasn't changed much, and has now become the *de-facto* standard for low-power wireless radio chips. A majority of low-power wireless standards build on top of IEEE802.15.4.

One major constraint of IEEE802.15.4 is that the size of its Protocol Data Unit (PDU) is limited to 127 bytes. This is small compared to the classical 1500-byte PDU of Ethernet — a popular wired technology — and smaller than 1280 bytes, which is the minimal value expected by IPv6 for the Maximum Transmission Unit (MTU) of any given link. In 2007 the Internet Engineer Task Force (IETF) — the standardization body behind most protocols used in the Internet today — created the 6LoWPAN WG, which defined a compaction and fragmentation mechanism to efficiently transport IPv6 packets in IEEE802.15.4 frames.[2] This simple mechanism allows low-power wireless devices to appear as regular Internet hosts, thereby becoming the "fingers of the Internet." The Internet of Things (IoT) revolution had begun.

Driven by the endless possibilities of connecting "Things" to the Internet, several IETF working groups were created. They have standardized the IPv6 Routing Protocol for Low power and Lossy Networks (RPL) [4], a routing protocol designed to extend the range of those networks, and the Constrained Application Protocol (CoAP) [5], a protocol enabling client-server interaction between low-power devices and traditional Internet hosts. In parallel, a *detnet* effort is starting at the IETF to bring deterministic networking properties at both Layer-2 and Layer-3 in a homogeneous manner, extending the work that has started at the IEEE at 802.1 AVB (for audio/video bridging) in collaboration with 802.1 TSN (for time-sensitive networking).

While IPv6 capability significantly simplifies the integration of a low-power wireless network into a production system, it does not answer the main requirement: reliability. The wireless medium is unreliable in nature. A wireless link connecting two devices typically has an associated Packet Delivery Ratio (PDR) quantifying what portion of transmitted frames are received. This PDR depends largely on the environment surrounding the communication devices (walls, machinery, etc.). Since the environment changes, the PDR is not predictable in any practical sense. Two wireless phenomena severely impact the PDR: external interference and multi-path fading.

External interference is caused by a different technology (or an independent deployment of the same technology) impacting the wireless signals. Although some frequency bands are more crowded by different technologies than others, interference happens in all frequency bands, and using a dedicated frequency band, at best, pushes the problem further away.

Multi-path fading is less often taken into account, yet is in some sense more destructive than interference. It happens when multiple "echoes" of the same wireless signal interfere constructively and/or destructively at the receiver's antenna. These echoes have bounced off objects in the environment, and a slight change in the environment causes the PDR of a wireless link to severely swing, leading to very unstable connectivity within a network operating on a single frequency. Multi-path fading is extremely sensitive to both frequency and location, and will change dramatically if a device, or some reflector in the environment, moves.

Fortunately, frequency diversity is efficient at combating both phenomena, and is therefore used in many wireless technologies, from Bluetooth to cellular systems. Using multiple frequencies reduces the impact of interference, as interference typically affects only some of the available frequencies at a given moment. Also, different frequencies have different multi-path constructive/destructive self-interference patterns, so using multiple frequencies "smoothens" away the impact of multi-path fading.

In an IEEE802.15.4 system, frequency diversity can be obtained without changes in the physical layer through a technique known as "channel hopping." When channel hopping, sender and receiver devices change frequency at each packet transmission, following a pseudo-random hopping pattern. The result is that, if a transmission is unsuccessful (possibly due to external interference or multi-path fading), retransmission occurs on a different frequency. The key is that retransmitting on a different frequency has a higher probability of success than using the same frequency again.

The idea of channel hopping was a key enabler for industrial wireless, with now three protocols competing in the industrial process

While IPv6 capability significantly simplifies the integration of a low-power wireless network into a production system, it does not answer the main requirement: reliability. TSCH technology achieves over 99.999% end-to-end reliability and ultra-low power consumption.

[2] 6LoWPAN is now succeeded at the IETF by the 6lo working group, which maintains the protocols and generalizes the work to other media.

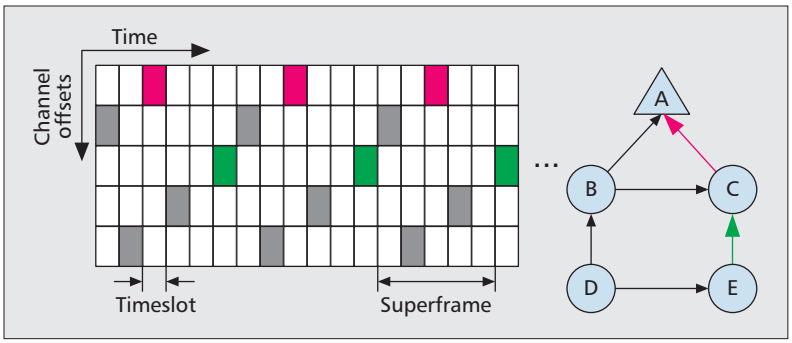


Figure 1. In a TSCH network, a schedule orchestrates all communication.

control space alone, all deriving from that same base technology.

The Emergence of Industrial Wireless

Channel Hopping is used in a class of communication protocols known as Time Synchronized Channel Hopping (TSCH). In a TSCH network, time is sliced into timeslots, and timeslots are grouped into a slotframe (typically 10's to 1,000's of timeslots long). The timeslot duration (typically 10ms) is large enough to accommodate the longest data frame, and leave time for the receiver to send back an acknowledgement (ACK) indicating successful reception. If the transmitter does not receive an ACK after transmitting, it can decide to re-transmit at a later time.

A schedule coordinates all of the communication in a TSCH network. It indicates to a node what do to in each timeslot: transmit, receive, or sleep. For each transmit or receive slot, the schedule also indicates the neighbor to communicate with, and a channel offset to communicate on. This channel offset is translated into a frequency on-the-fly using a pseudo-random pattern, resulting in channel hopping.

Timeslots in the schedule can be seen as atomic link-layer resources. Scheduling more transmit/receive timeslots increases the throughput of the network, but also increases the average power consumption of the nodes.

Figure 1 shows a canonical example of a communication schedule. Time is sliced up into timeslot, and (in this example) four timeslots form a slotframe that repeats over time. Each row represents a different channel offset; multiple communications can happen in the network at the same time, but on a different channel offset without interfering. When node *E* wants to send a packet to node *A*, it waits for cell colored green to send the packet to *C*. Similarly, node *C* sends it to *A*. If any of those transmissions fail (i.e. the transmitter does not receive an ACK), the transmitter tries again at the next opportunity, possibly at the next iteration of the slotframe.

Channel hopping, which exploits frequency diversity, can be combined with path diversity. With the latter, a packet can travel multiple disjoint paths to the destination, giving additional redundancy to further increase the reliability of the network.

TSCH technology was first commercialized by Dust Networks as Time Synchronized Mesh Protocol (TSMP), which achieves over 99.999% end-to-end reliability and ultra-low power consumption. Elements of this technology were adopted in WirelessHART [6] (IEC62591). TSCH is also at the heart of ISA100.11a (IEC 62734). TSCH has become the *de-facto* technology for highly reliable, low-power wireless sensor networking technology, with tens of thousands of networks deployed today.[3]

In 2012 the IEEE802.15.4e-2012 amendment [7] was published, reusing the core ideas of TSCH in a well-layered approach, allowing "upper stack" protocols to run on top.

For the IEEE802.15.4e TSCH with IETF standards such as 6LoWPAN, RPL and CoAP combination to work, a standardization "gap" needs to be filled. IEEE802.15.4e TSCH defines what a node does to execute a schedule, but does not detail how to build and maintain that schedule. Similarly, an IETF standard such as RPL organizes an existing topology into a multihop routing structure, but is agnostic to the underlying link layer technology, and hence to the notion of a TSCH communication schedule.

What is missing is a sublayer that allows a scheduling entity to manage the TSCH schedule in the network. This standardization "gap" is currently being filled by 6TiSCH ("IPv6 over the TSCH mode of IEEE 802.15.4e"), a new IETF working group dedicated to enabling IPv6 over the TSCH mode of the IEEE802.15.4e standard.[4] The resulting protocol stack is depicted in Fig. 2. 6TiSCH, and the scheduling mechanisms it introduces, are detailed below.

6TiSCH: IPv6 over IEEE802.15.4e TSCH

In a TSCH network, the communication schedule orchestrates all communication. Building a communication schedule involves assigning timeslots to communication between neighbor nodes, as atomic link-layer resources. Assigning multiple timeslots to the same neighbors increases the available throughput (the number of packets these neighbors can exchange per second), and lowers the latency of that communication. It also requires the radios of those nodes to be *on* more often, thereby increasing the average energy consumption, resulting in a shorter battery lifetime.

6TiSCH defines a new global concept that is called a Channel distribution/usage (CDU) matrix; a Channel distribution/usage (CDU) matrix is a matrix of so-called "cells" with a height equal to the number of available frequencies (indexed by ChannelOffsets), and a width in timeslots that is the period of the network scheduling operation (indexed by slotOffsets). The CDU matrix can be partitioned into chunks. As seen in Fig. 3, a Chunk is a well known list of cells, well distributed in time and frequency, within the CDU matrix. The partition of the CDU in chunks is globally known by all the nodes in the network to support the appropriation process, which is a negotiation between nodes within an interference domain. As illustrated in Fig. 4, a node that appropriates a chunk gets to decide which transmissions will occur over the cells in the chunk within its interference domain. Ultimately, a chunk represents some amount of bandwidth and can be seen as the generalization of a transmission channel in the time/frequency domain.

[3] Including over 35,000 networks currently operating Dust Networks' TSCH-based SmartMesh products [2].

[4] http://tools.ietf.org/wg/6tisch/charters.

Architecture

6TiSCH is chartered with the uncommon milestone of delivering an architecture that encompasses multiple standards from multiple areas within the IETF, as well as external work from Standards Development Organizations (SDOs) such as the IEEE. The 6TiSCH architecture is currently being designed with a goal to provide the high delivery ratios and deterministic jitter and latency that are typical to Industrial Wireless Sensor Network (WSN) protocols such as WirelessHART and ISA100.11a, and at the same time also enable large-scale monitoring deployments in order to achieve the Industrial Internet paradigm.

In order to scale while maintaining throughput, a 6TiSCH network is envisioned to typically comprise a number of small wireless mesh networks federated by a spanning high-speed backbone. The 6LoWPAN Neighbor Discovery (ND) operation is leveraged inside the meshes to eliminate the need for widespread multicast inherent to the classical IPv6 ND operation, and a new functionality is introduced in the backbone router to redistribute that operation inside the backbone, effectively forming a large IPv6 multilink subnet, as illustrated in Fig. 5.

The experience from industrial WSN supports a centralized approach, enabling a total control and optimization of the network operation from a central compute engine with a "God's view," which is called a Network Manager or a System Manager in the art, and which corresponds to the IETF concept of a Path Computation Element (PCE). The PCE considers all the flows and all the network capacity and computes the optimum set of end-to-end tracks, which are pushed to the network to support individual flows in a pre-determined fashion. In a number of aspects, the centralized approach simplifies the problem: the number of nodes is limited in a given mesh to limit the computational complexity, and all the nodes are expected to be in a same interference domain, so any given slot is attributed at most to one critical flow. In that model, each flow is associated with a reserved track, leading to a deterministic use of the medium.

On the other hand, the experience from the Advanced Metering Infrastructure (AMI)/Automatic Meter Reading (AMR) space demonstrates the value of distributed routing with RPL, and when applicable, distributed scheduling. RPL, the IPv6 Routing Protocol for low-power wireless networks, is the 4th routing protocol standardized by the IETF, after RIP, OSPF, and BGP. RPL is a generic distance-vector (DV) protocol that was designed at the IETF to be very economical in the control plane so as to serve Internet of Things (IoT) applications. The protocol computes abstract Directed Acyclic Graphs (DAG) to support Non-Equal Cost Multipath (NECM) redundant topologies that are optimized for various application needs by specific Objective Functions, which are a kind of plug-in to the generic protocol. In that model, multiple flows are merged along a DAG, leading a statistical use of the medium.

TSCH has major differences with Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA), widely used in wireless technologies. The low-power nature of TSCH allows routing nodes to *not* need line power, opening up many new applications. Moreover, with TSCH, the spectrum is partitioned in timeslots that are individually reserved to one communication, whereas with CSMA every device may get access to the spectrum at any time. The reservation mechanism is required to protect the most critical flows, but yields the drawback of potentially wasting resources when the critical flows are not present. In order to compensate, the 6TiSCH architecture supports the opportunistic reuse of timeslots along deterministic tracks in the centralized approach. In the distributed approach, the effective use of the bundles of timeslots that implement the link abstraction for the IP routing function is much more dynamic, and avoiding waste is much more challenging. Mechanisms such as On-the-Fly scheduling are evaluated to provide a most reactive arbitration with a very limited signaling overhead, as detailed later.

The 6TiSCH architecture must also cover all-encompassing components such as security and management. These components affect many aspects of the device operations, and for each, several techniques exist in the art that could be relevant for 6TiSCH. But the constraints in memory, CPU, bandwidth, and latency render most existing solutions impractical, and the need for high scalability with low maintenance encourages the use of more bleeding edge autonomic practices. 6TiSCH aims at reusing code for multiple components so as to avoid any functional duplication. This is why the group is considering management over CoAP so as to reuse the code that will serve to report sensor data, and the Datagram Transport Layer Security (DTLS) security that protects it. For the same reason, the group is also considering using Extensible Authentication Protocol (EAP) — Transport Layer Security (TLS) to protect the device communications from the

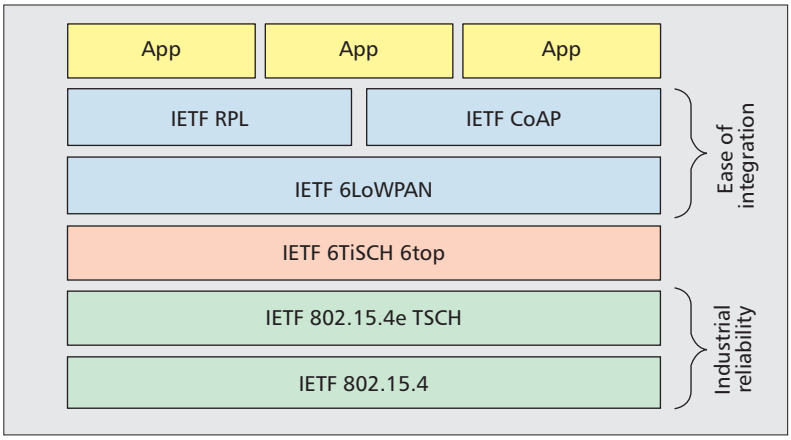


Figure 2. The protocol stack for the Industrial Internet (of Things).

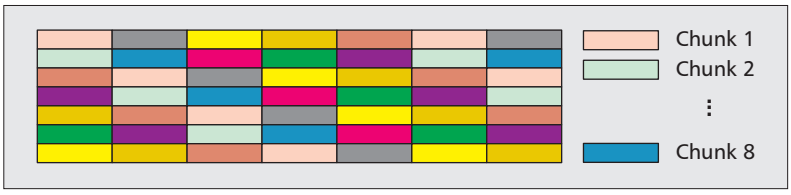


Figure 3. Example of chunks in a channel distribution/usage matrix (CDU).

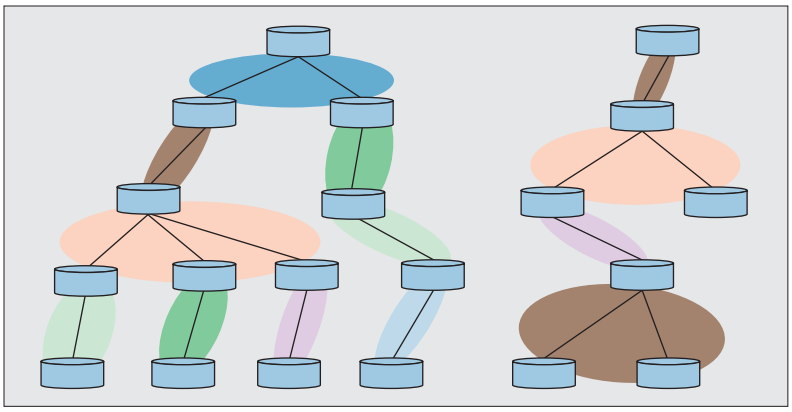


Figure 4. Chunk assignment for an example node topology.

joining time on, sharing capabilities to parse certificates and support public keys between the authentication and authorization flows in the join process and the data traffic. The support of a shared secret to bootstrap a network may be acceptable on some deployments that are limited in number, space, and time, but is not generally a recommended idea, and though it is not barred by the architecture, it will not be documented either.

Scheduling Mechanisms

Scheduling a TSCH network involves a policy using several mechanisms to manage the TSCH schedule operating in the network.

The policy is in charge of determining which timeslot is allocated to which nodes. Scheduling can be seen as an optimization problem where timeslots are assigned to pairs of communicating neighbor nodes to satisfy application-level networking requirements. These requirements can be expressed as metrics that the scheduling algorithm needs to optimize. Metrics can involve allocation collisions, energy consumption, latency, or a combination thereof. The scheduling algorithm itself can follow a variety of approaches, including genetic algorithms, stochastic approaches, or combinatorial optimization. To allow maximum flexibility, the scheduling policy is *not* standardized by 6TiSCH.

6TiSCH makes no assumption of the scheduling entity. It can be centralized (an entity in the network, for example the PCE in Fig. 5, is in charge of computing a schedule) or distributed (nodes communicate with one another to agree on a schedule).

The scheduling entity relies on scheduling mechanisms (e.g. packet formats) to access the TSCH schedule on the nodes in the network, thereby controlling the communication schedule operating in the network. 6TiSCH is standardizing these mechanisms to support the different scheduling approaches detailed in the following paragraphs.

In the simplest case, the TSCH schedule can be static. Once a node has acquired this schedule (it is either pre-programmed, or learned during the joining process), the schedule is never changed. Since no scheduling entity is required in this case, this "minimal" schedule [8] can be used as a bootstrap mechanism (i.e. operating before a scheduling entity is operational) or a fall-back mechanism (i.e. operating after the scheduling entity has failed). The pre-configured slots can be seen as a control plane to enable other mechanisms to operate the network.

Another approach is centralized scheduling. 6TiSCH is standardizing a set of CoAP resources [9] enabling an external entity to control the TSCH schedule. These CoAP resources turn a node into the constrained equivalent of a web server. The central scheduler acts as a browser, downloading current state and uploading new configurations to that CoAP server.

Another alternative is distributed scheduling, in which neighbor nodes negotiate the use of timeslots with one another, without requiring intervention from a central entity. 6TiSCH is standardizing the format of the packets exchanged between neighbor nodes to conduct this negotiation [10].

The distributed scheduling policy currently being finalized by 6TiSCH is called "On-The-Fly scheduling" (OTF) [11]. When using OTF, a node monitors the state of its outgoing packet queue (stored in its internal memory). If the queue fills up, OTF determines that there is not enough "outbound" bandwidth, and triggers the negotiation of additional timeslots with the appropriate neighbor(s). Similarly, if the queue is often empty, it negotiates the removal of outbound timeslots. OTF describes the structure, policies, and interfaces of the distributed scheduling scheme, while leaving the bandwidth estimation algorithm out-of-scope for greater flexibility.

Open Issues and Outlook

In accordance with its chartered mission to document an overall architecture, 6TiSCH needs to examine building blocks that were designed in different working groups at the IETF and that are not necessarily optimized to fit with one another, and then to steer work in the original working group(s) so as to round up the angles.

A typical example of this is the need to enable some interaction between 6LoWPAN-ND [12] and the RPL routing protocol [4]. 6LoWPAN-ND is an extension to IPv6 Neighbor Discovery [13] that is adapted to low-power, duty-cycled devices. 6LoWPAN-ND replaces the classical ND model (heavily based on multicast, which can be inefficient in low-power wireless systems) with a registration model involving only unicast communication. RPL was specifically designed to address low-power wireless networks and constrained devices, but the information in 6LoWPAN-ND is not sufficient for a RPL router to represent a 6LoWPAN node adequately.

The most intriguing aspect of the work to come is in the distribution of schedule and more specifically in the dynamic allocation of time slots in the distributed approach. Basically, the question is how to resolve the tension between the optimal usage of the bandwidth for statistically multiplexed traffic and the reservation mechanism that is inherent to exclusive timeslotted operation.

Conclusion

The Internet Protocol (IP) is the cornerstone of today's Internet. Through an important standardization effort at the IETF, standards such as 6LoWPAN allow low-power wireless devices to

behave like any other Internet host, greatly simplifying the integration of low-power wireless networks into a larger networking system.

Low-power wireless technologies based on Time Synchronized Channel Hopping (TSCH) proved to satisfy the stringent reliability and low-power requirements of industrial applications, and therefore become part of the heart of standards such as WirlessHART, ISA100.11a, and IEEE802.15.4e TSCH.

These standardization efforts triggered the trend of industrial deterministic networks to both "go wireless" and "go IP." For this convergence to be possible, however, a standardization gap had to be filled: standards are needed to allow TSCH schedules to be managed in an IP-enabled infrastructure, thus empowering industrial performance with the ease-of-use of IP.

Closing that standardization gap is the goal of the newly created 6TiSCH working group ("IPv6 over the TSCH mode of IEEE 802.15.4e") at the IETF, dedicated to enabling IPv6 over the TSCH mode of the IEEE802.15.4e standard. 6TiSCH is standardizing mechanisms supporting different scheduling approaches, including centralized, distributed, and hybrid.

With this effort, industrial applications are heading toward the integration of Information and Operation Technologies. The use of a common protocol stack to enable seamless communication between heterogeneous devices, from powerful data servers to tiny sensing nodes, has the potential to create a wide range of applications, including those based on Big Data storage and analysis engines.

Acknowledgments

This work was supported by Anillo Project ACT-53, Fondecyt project No. 11121475, CIRIC (INRIA-Chile) Project "Network Design," Project Semilla — UDP "ANDES" and "Análisis y diseño de algoritmos en redes de bajo consumo aplicado a condiciones extremas de los Andes": Programa de Cooperación Científica Internacional CONICYT/MINCYT 2011.

References

[1] H. C. Foundation, "HART Communication Protocol Specification," HART Communication Foundation Std. HCF_SPEC-13, Rev. 7.5, 29 May 2013.
[2] T. Watteyne *et al.*, "Technical Overview of SmartMesh IP," *Int'l. Wksp. Extending Seamlessly to the Internet of Things (esIoT)*, Taiwan, 3-5 July 2013.
[3] "802.15.4-2011: IEEE Standard for Local and metropolitan area networks — Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs)," IEEE Std., 5 Sept. 2011.
[4] T. Winter *et al.*, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," IETF Std. RFC6550, Mar. 2012.
[5] Z. Shelby, K. Hartke, and C. Bormann, "Constrained Application Protocol (CoAP)," Internet-Draft [work-in-progress], IETF Std., Rev. draft-ietf-core-coap-18, 28 June 2013.
[6] "WirelessHART Specification 75: TDMA Data-Link Layer," HART Communication Foundation Std., Rev. 1.1, 2008, hCF_SPEC-75.
[7] "802.15.4e-2012: IEEE Standard for Local and Metropolitan Area Networks — Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 1: MAC Sublayer," IEEE Std., 16 Apr. 2012.
[8] X. Vilajosana and K. Pister, "Minimal 6TiSCH Configuration," Internet-Draft [work-in-progress], IETF Std., Rev. draft-ietf-6tisch-minimal-00, 19 Nov. 2013.
[9] R. Sudhaakar and P. Zand, "6TiSCH Resource Management and Interaction using CoAP," Internet-Draft [work-in-progress], IETF Std., Rev. draft-ietf-6tisch-coap-00, 6 May 2014.
[10] Q. Wang, X. Vilajosana, and T. Watteyne, "6TiSCH Operation Sublayer (6top)," Internet-Draft [work-in-progress], IETF Std., Rev. draft-wang-6tisch-6top-sublayer-00, 10 Apr. 2014.
[11] D. Dujovne *et al.*, "6TiSCH On-the-Fly Scheduling," Internet-Draft [work-in-progress], IETF Std., Rev. draft-dujovne-6tisch-on-the-fly-02, 14 Feb. 2014.

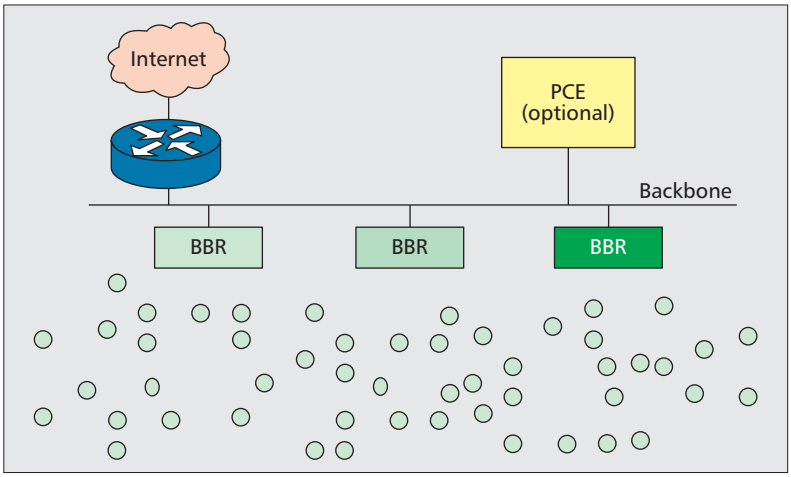


Figure 5. Architecture of a full-featured 6TiSCH network, where multiple backbone routers interconnect different mesh networks.

[12] Z. Shelby *et al.*, "Neighbor Discovery Optimization for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs)," IETF 6LoWPAN Std. RFC6775, Nov. 2012.
[13] T. Narten *et al.*, "Neighbor Discovery for IP version 6 (IPv6)," IETF Std. RFC4861, Sept. 2007.

Biographies

Diego Dujovne (diego.dujovne@mail.udp.cl) obtained his electronic engineering degree from the Universidad Nacional de Córdoba (UNC), Argentina, in 1999, developed his Ph.D. at INRIA Sophia Antipolis at Équipe-Projet PLANETE, and obtained his degree from UNSA, France, in 2009. For five years he developed several university-industry collaboration projects on instrumentation and communications at LIADE, UNC. He is currently a full-time academic at the Universidad Diego Portales, Chile. His current research interests include IPv6 over LLN routing and scheduling, and wireless experimental platform development and measurement methodologies.

Thomas Watteyne (twatteyne@linear.com) is a senior networking design engineer at Linear Technology, in the Dust Networks product group, which specializes in ultra-low power and highly reliable wireless sensor networking. He designs networking solutions based on a variety of IoT standards, and promotes the use of highly reliable standards such as IEEE802.15.4e. He is co-chairing the new IETF 6TiSCH working group, which aims at standardizing how to use IEEE802.15.4e TSCH in IPv6-enabled mesh networks. Prior to Dust Network, Thomas was a postdoctoral researcher at the University of California, Berkeley, working with Prof. Kristofer Pister. He started Berkeley's OpenWSN project, an open-source initiative to promote the use of fully standards-based protocol stacks in M2M applications. Between 2005 and 2008 he was a research engineer at France Telecom, Orange Labs. He obtained his Ph.D. in computer science (2008) and MSc in telecommunications (2005) from INSA Lyon, France.

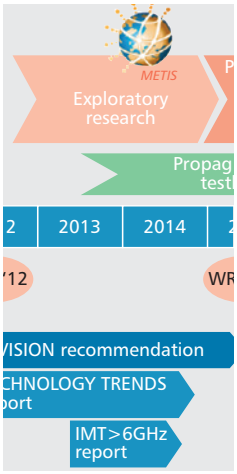
Xavier Vilajosana (xvilajosana@worldsensing.com) is an entrepreneur and co-founder of Worldsensing. He is an associate professor at the Open University of Catalonia (UOC). From January 2012 to January 2014 he was a visiting professor at UC Berkeley holding a prestigious Fulbright fellowship. In 2008 he was a researcher at France Telecom R&D Labs, Paris. He is one of the main promoters of low power wireless technologies, co-leading the OpenWSN.org initiative at UC Berkeley, and promoting the use of low power wireless standards for the emerging Industrial Internet paradigm. He is also author of different RFCs, as part of his standardization activities for low power industrial networks. He has an M.Sc. degree in computer sciences from the Universitat Politécnica de Catalunya (UPC) and a Ph.D. in computer science from the UOC. He currently holds eight patents and more than 20 high impact journal publications.

Pascal Thubert (pthubert@cisco.com) has been involved in research, development, and standards efforts for evolving Internet and wireless technologies since joining Cisco in 2000. He currently works within Cisco's Chief Technology and Architecture Office (CTAO), where he focuses on industrial and other deterministic networks and products in the general context of the Internet of Everything. He is co-leading 6TiSCH, the IETF standard for IPv6 over the 802.15.4e TSCH deterministic MAC, and DetNet, a new joined effort with IEEE for cross layer deterministic networks. Earlier, he specialized in IPv6 as applied to mobility and wireless devices and worked in Cisco's core IPv6 product development group. In parallel with his R&D missions, he has authored multiple IETF RFCs and draft standards dealing with IPv6, mobility, and the Internet of Things. In particular, he participated as co-editor of the ISA100.11a specification, as well as the NEMO, 6LoWPAN and RPL IETF standards, and participated actively in the introduction of the IT/OT convergence for an Industrial Internet.

5G Wireless Access: Requirements and Realization

We are just at the beginning of a transition into a fully connected Networked Society that will provide access to information and sharing of data anywhere and anytime for anyone and anything. Thus, in the future wireless access will not only be about connectivity for people but for anything that benefits from being connected.

Erik Dahlman, Gunnar Mildh, Stefan Parkvall, Janne Peisa, Joachim Sachs, Yngve Selén, and Johan Sköld



Abstract

5G, the mobile communication technology for beyond 2020, will provide access to information and the sharing of data anywhere and anytime for anyone and anything. This paper describes the current status of the processes moving toward 5G, or "IMT for 2020 and beyond," in ITU-R. We also provide a view of 5G opportunities, challenges, requirements and technical solutions.

Introduction and Challenges

Mobile communication systems have evolved from supporting analog voice only to powerful systems providing hundreds of thousands of different applications to billions of users. We are just at the beginning of a transition into a fully connected Networked Society that will provide access to information and sharing of data *anywhere* and *anytime* for *anyone* and *anything*. Thus, in the future wireless access will not only be about connectivity for people but for anything that benefits from being connected. This includes such diverse things as household appliances, traffic control and safety functions, infrastructure monitoring systems, medical equipment, and much more. As a consequence, compared to the wireless networks of today, next-generation wireless access will support a much wider range of use case characteristics and corresponding access requirements.

Next-generation mobile communication will not be available until after 2020. The global research efforts already underway are exemplified by [1], as well as by projects such as the METIS project in Europe [2]. Research is likely to continue for a few more years from now before the standardization and eventual commercialization of the system begin. However, we already have a relatively clear view of the main challenges and opportunities, as well as the key technology components of the future 5G systems [3].

Conventional mobile-broadband (MBB) applications will continue to drive demand for higher traffic capacity and higher end-user data rates within the wireless-access network. In terms of traffic demand, predictions range from hundreds of times to more than a thousand times higher traffic in the next 10 years [3, 4]. Most of this traffic, primarily video, will come from "conventional" mobile broadband access. But compared to the networks of today future wireless networks must offer radically lower cost and energy consumption per delivered bit to carry the massive traffic affordably and sustainably.

The provisioning of higher end-user data rates, allowing for faster access to information, has been the key driving force for the development and evolution of 3G and 4G wireless-access technologies. This quest will continue in the future. For next-generation wireless-access networks we envision data rates on the order of 10 Gb/s for specific scenarios, such as indoor offices and university campuses. More importantly, data rates of more than 100 Mb/s should be generally available in urban and suburban environments. And finally, to provide a truly ubiquitous connectivity, rates of at least a few Mb/s should be available essentially everywhere, including far-off rural and deep indoor environments.

However, enhanced mobile broadband with its corresponding demand for higher traffic capacity and data rates will be only one of the drivers for the development of next-generation wireless access. In addition, new machine-type-communication (MTC) use cases will impose other sets of requirements, as shown in Fig. 1 and described in the following.

Although the access latency offered by LTE is sufficient for most mobile-broadband applications, it may not be sufficient for latency-critical applications, such as traffic safety, infrastructure protection, or emerging industrial Internet applications. To ensure support for such *mission-critical* MTC applications, next-generation wireless access should allow for latencies on the order of 1 ms or less.

Even more important, full support for mission-critical MTC applications will require ultra-reliable connectivity with essentially guaranteed availability. Also, reliability-of-service will have to be orders of magnitude higher than current, already highly reliable, networks.

On the other side of the scale is a vision of *massive* MTC connectivity, with tens of billions of low-cost connected devices and sensors. To realize this vision, 5G must enable the availability of truly low-cost devices. Furthermore, if they are to operate over several years without recharging, future wireless devices, including all sorts of sensors, should operate with extremely low energy consumption.

Thus, as outlined in Fig. 2, next-generation wireless access should extend the performance and capability of wireless access networks in many dimensions. As indicated in the figure, some of these extensions and enhancements — for example, in terms of traffic capacity and high data rates — are primarily driven by the conventional MBB use case. Others, requiring extreme reliability and support for truly massive numbers of devices, are driven more by new use cases, for example, mission-critical and massive MTC.

The authors are with Ericsson Research.

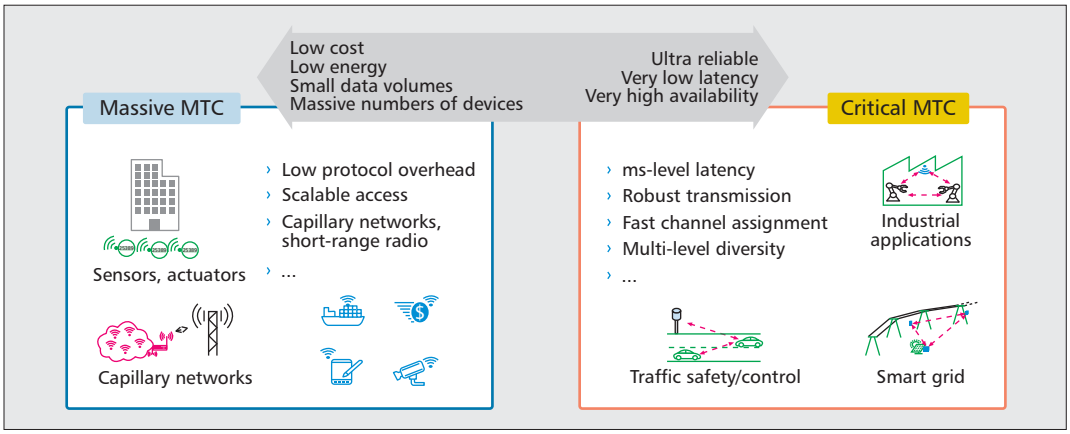


Figure 1. Novel machine-centric use cases for 5G and corresponding requirements.

The ITU-R Process for Defining 5G

ITU-R Working Party (WP) 5D is responsible for the overall radio system aspects of International Mobile Telecommunications (IMT) systems. These consists at present of IMT-2000 (3G) and IMT-Advanced (4G). WP5D has the prime responsibility within ITU-R for issues related to the terrestrial component of IMT, including technical, operational and spectrum-related issues. For this purpose, the group develops and maintains recommendations, including the radio interface specifications for IMT-2000 systems in ITU-R Recommendation M.1457 [5] and for IMT-Advanced in ITU-R Recommendation M.2012 [6].

WP5D is now working on three deliverables for "IMT for 2020 and beyond," which corresponds to 5G, for the next World Radio Congress WRC-15. The main deliverable is a recommendation referred to as the *IMT Vision* [7]. It will define the:

- Roles of IMT in future society.
- Framework and overall objectives of the future development of IMT for 2020 and beyond.
- Key capabilities and technical enablers.

This recommendation will be complemented by two technical reports. One report on *technology trends* [8] will describe the technical aspects of terrestrial IMT systems during 2015-2020 and beyond, including the evolution of IMT. A second report will describe the technical feasibility of IMT in the bands above 6 GHz [9].

As shown in Fig. 3, the recommendation for the IMT Vision will be completed in advance of WRC-15; the above mentioned two reports will then serve as input for completing the recommendation. This initial work within ITU-R ties in well with the ongoing global exploratory research activities on 5G in METIS and other projects. The next steps after WRC-15 are for WP5D to set technical performance requirements and to develop an evaluation process for 5G. Following the evaluation, a consensus-building process will result in radio interface specifications for 5G, similar to the ones for IMT-2000 and IMT-Advanced [5, 6]. Note that so far no official name has been chosen for 5G in ITU-R.

An essential part of the IMT Vision is to define new key capabilities for 5G, and to relate these to the key use cases and scenarios for 5G. Many key capabilities are enhancements of existing 3G and 4G capabilities, while others are new capabilities not originally envisioned for 3G and 4G. Figure 2, currently under discussion in WP5D, shows one way of illustrating capabilities deemed essential for 5G. It further connects these key capabilities to the main use cases driving them.

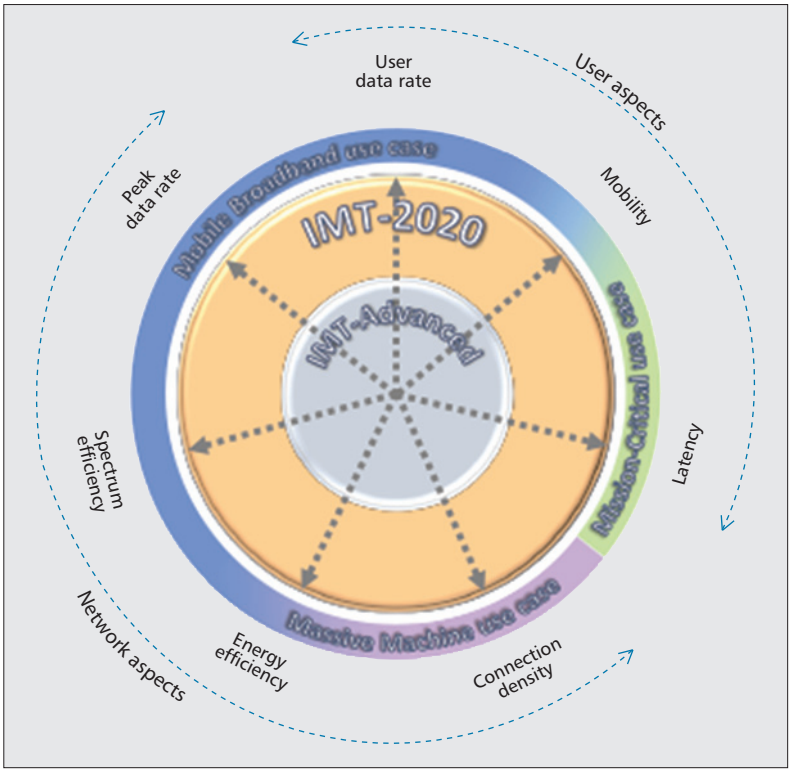


Figure 2. 5G use-case categories and the corresponding key areas for enhancements.

Spectrum for 5G

To enable the expected massive traffic increase, additional spectrum will have to be assigned to mobile wireless communications. For the 2015 World Radio Conference (WRC-15), the focus

To fulfill long-term traffic demands and, perhaps even more important, to enable the very wide transmission bandwidths needed for multi-Gb/s data rates efficiently, next-generation wireless access will extend the range of operation to frequencies above 10 GHz.

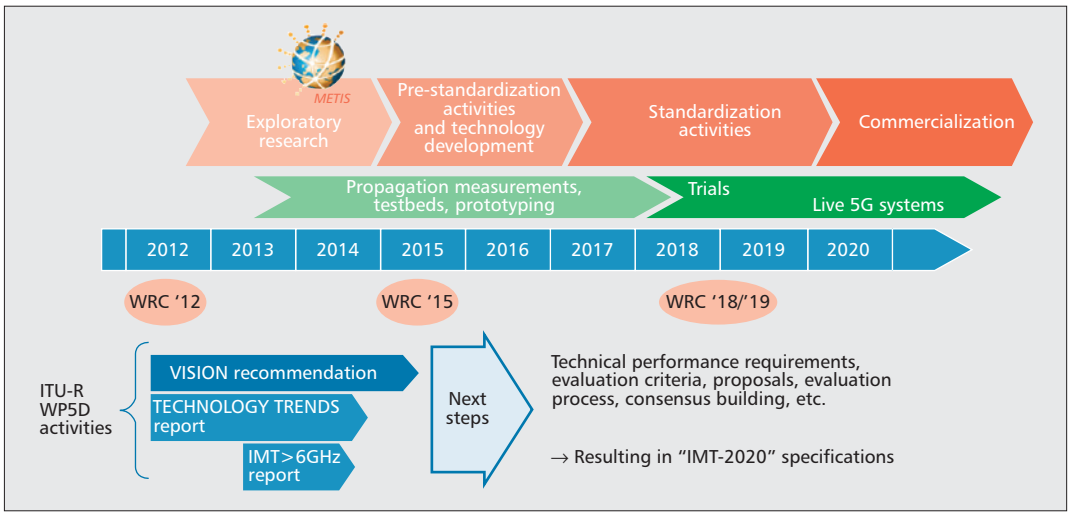


Figure 3. ITU-R Time plan for the work on IMT for 2020 and beyond.

will be on ensuring additional spectrum below 6.5 GHz. However, to fulfill long-term traffic demands and, perhaps even more important, to efficiently enable the very wide transmission bandwidths needed for multi-Gb/s data rates, next-generation wireless access will extend the range of operation to frequencies above 10 GHz.

Identifying and allocating new spectrum beyond 10 GHz for mobile wireless communications is expected to be on the agenda at WRC-18/19. At this stage the entire range of spectrum from 10 GHz up to 70 GHz or even higher (well into what is commonly referred to as the millimeter-wave [mmW] range) is being considered. Research and concept development of future wireless access must cover the entire spectrum, from currently used spectrum at 1 GHz and even lower all the way up to and including mmW frequency bands.

We do not believe that addressing this very wide range of frequencies with a single radio-interface structure is the best approach. Issues such as propagation characteristics, implementation aspects, and, for certain frequency bands, compatibility with legacy technologies impose different constraints that affect the basic radio-interface design (see Fig. 4).

Up to a certain frequency range, radio interface design can be based on the same principles as current wireless technologies. One would then assume relatively wide-area coverage, high-performance radio-frequency (RF) design, and so on. An OFDM-based transmission technology will most likely remain a good baseline, although the detailed numerology would probably need to be adjusted to match frequencies above 10 GHz.

However, for even higher frequency bands, propagation characteristics and implementation aspects speak in favor of a more simplified radio-interface structure targeting shorter range communications (so called ultra-dense deployments), allowing for more relaxed requirements on, for instance, RF parts.

Furthermore, one needs to take into account that around 2020, when next-generation wireless access is expected to reach the market, LTE will be heavily deployed in licensed spectrum below 6.5 GHz. It is highly desirable if next-generation wireless access can be introduced without impacting existing deployments and that service can be provided to existing user devices. Thus, in such frequency bands, it should be possible to introduce next-generation wireless functionality while retaining compatibility with existing technology, primarily LTE.

Technical Solutions

Below we discuss some key technology solutions that we believe will be important components for future 5G wireless access.

Massive Beamforming and Advanced Antennas

Advanced antennas with multiple antenna elements can improve coverage for high-data-rate communications as well as significantly increase overall system capacity. Beamforming, where multiple antenna elements are used to form narrow beams, is an efficient tool for improving both data rates and capacity. Spatial multiplexing, where propagation properties are exploited to provide multiple data streams simultaneously to one or more terminals, is another example of an important multi-antenna technique.

To some extent, these techniques are integral parts of LTE, but their full potential remains to be unleashed when they play an even bigger role in future systems. At higher frequency bands, propagation conditions are more challenging than on current LTE bands. Higher diffraction and outdoor-to-indoor losses lead to a correspondingly more challenging link budget. The output power of the equipment — in particular, the mobile terminals — may because of regulatory restrictions also be more limited than at lower frequency bands. Therefore, extensive use of beamforming, in particular at the base station, is an essential part of high-frequency wireless access. The challenging propagation conditions also call for dense network deployments, implying that networks operating in high-frequency bands will be primarily deployed in densely populated areas, such as city centers, airports, train stations, and indoor offices.

Advanced antennas with massive numbers of elements (known as massive MIMO) can also be used to reduce the impact of RF imperfections and to control the interference distribution in the network.

Ultra-Lean Design

Current cellular systems continuously transmit reference signals and broadcast system information that is used by terminals as they move across cells. With denser deployment and more network nodes, such "always-on" transmissions are not attractive from an interference and energy consumption perspective. Furthermore, in a heterogeneous deployment the overlaid macro nodes can provide system information and mobility assistance, thereby reducing the amount of system-related transmissions from the underlaid nodes. Ultra-lean transmission with "always-on" signals reduced to a bare minimum should therefore be a key design principle for future systems. Not only does ultra-lean transmission result in a very energy-efficient network, which translates into lower operational cost, it also reduces the overall interference level in the network. This is a critical enabler for very dense local-area deployments because the end-user performance would otherwise be limited by interference at low-to-medium loads.

For wireless access in the higher frequency bands, where networks are yet to be deployed, ultra-lean design is essential. For the lower frequency bands where a relatively large number of terminals are already deployed, the same basic principles can be applied, although backward compatibility needs special attention.

Spectrum Flexibility

Traditionally, cellular systems are deployed in exclusively licensed spectrum. Such a licensing regime will continue to play a key role to control interference and guarantee coverage. However, especially at higher frequency bands, future systems should provide a higher degree of spectrum flexibility. Unlicensed spectrum can be used to boost capacity, preferably in combination with licensed spectrum for critical control signaling and handling of mobility. Licensed-shared access, where the cellular system can access additional spectrum otherwise reserved for other uses, is another example of spectrum flexibility.

Flexible duplex, in which spectrum resources are dynamically assigned to either transmission direction, allows up to the full bandwidth to be opportunistically used in each direction. Moreover, flexible duplex can easily exploit unpaired spectrum allocations, which are more likely for large amounts of contiguous spectrum. Full-duplex communications with *simultaneous* transmission and reception on the same carrier can also be used, assuming appropriate interference cancellation techniques are available. Full-duplex operation at the network-side only — that is, receiving from one terminal while simultaneously transmitting to another — may be an interesting alternative compared to requiring the terminal also to handle simultaneous reception and transmission.

Flexible duplex is particularly attractive for small cells with similar terminal and network transmission power and where the strict isolation between uplink and downlink across cell borders provided by current FDD and TDD deployments is less important. Initial steps in this direction have recently been taken in LTE Rel-12.

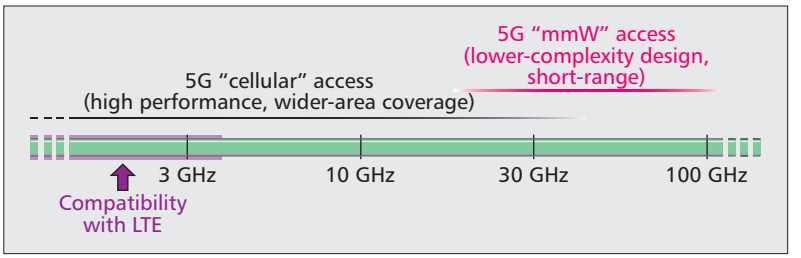


Figure 4. Spectrum range to be considered for 5G wireless access.

Low Latency

Lower latency over the radio link can be achieved by reducing transmission-time intervals and widening the bandwidth of radio resource blocks in which a specific amount of data is transmitted. This should be complemented by designing a physical-channel structure that allows for fast decoding at the receiver to reduce processing delays. To avoid queuing delays at the radio transmitter the medium-access control should be designed to enable immediate access. This can be achieved by providing instant-access resource allocations dimensioned to minimize collision risks. For some use cases, low-latency communications is required between devices in close proximity. In this case a direct device-to-device communication link can help providing low-latency transmission.

For some new use cases, like mission-critical MTC applications, a very high level of reliability of connectivity can be required, with low latencies provided with an extremely high level of certainty. Maintaining multiple connectivity links simultaneously can provide diversity and redundancy to address such stringent requirements.

Convergence of Access and Backhaul

In future systems, the traditional split between access and (wireless) backhaul links will likely diminish and the overall system design will not make a major distinction between the two. This brings several benefits. Wireless connectivity between radio network nodes and the rest of the network simplifies deployment, especially in a dense deployment with its large number of nodes. It is also an attractive alternative to deploying optical fiber, particularly at higher frequency bands with the availability of larger amounts of spectrum in combination with extensive beamforming and low-latency transmissions.

Wireless backhaul is in itself not new, but compared to the traditional fixed division of spectrum resources between access and backhaul links, spectrum resources are used more efficiently with a dynamic split between the two. This is facilitated by using the same radio interface technology for both types of links. Another benefit of this approach is that the same operational and maintenance systems can be used for both links.

It is desirable to achieve "zero-overhead" communications by simplifying connectivity states for devices and providing channel access with minimal signaling. Maximizing the devices' sleep opportunities can minimize energy consumption, leading to long battery life.

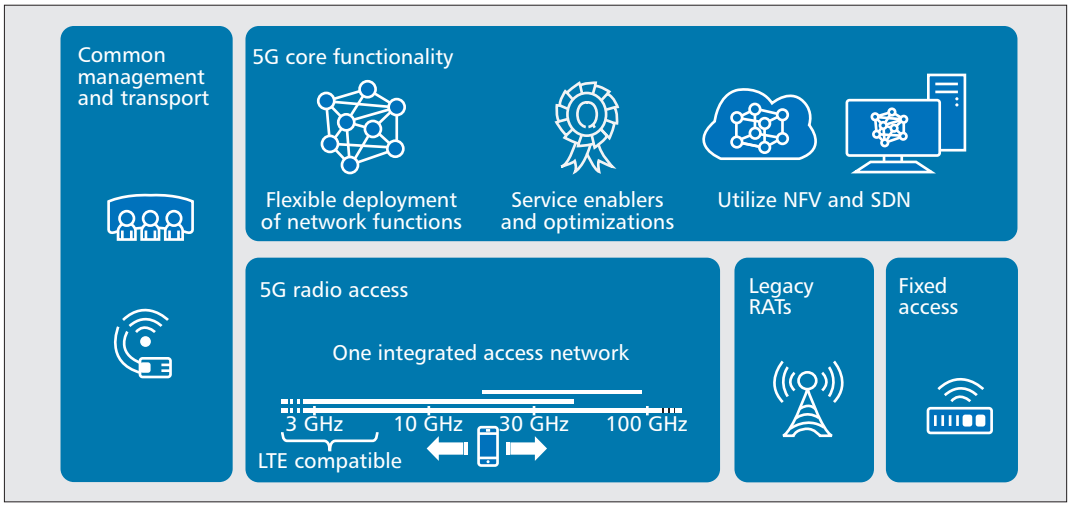


Figure 5. A visualization of the overall 5G architecture.

Enablers for Massive Machine-Type Communications

The transition to a Networked Society will lead to a massive number of connected devices, which transmit small amounts of data infrequently. These devices will often be simple and invisibly embedded into the fabric of the environment. This requires lightweight radio-module design and communication modes streamlined to the relaxed communication requirements. Devices should be able to operate for years on tiny batteries. It is desirable to achieve "zero-overhead" communications by simplifying connectivity states for devices and providing channel access with minimal signaling. Maximizing the devices' sleep opportunities can minimize energy consumption, leading to long battery life.

According to [10], we expect that it should be possible for a device to transmit 1 kbyte of data every 10 minutes and run for 10 years on a single AA battery. With connected devices in remote and challenging locations with severe path loss, optional transmission modes should provide connectivity at low rates, with control channels that provide the required robustness efficiently. Typically, massive machine-type communications will take place at frequency bands below 3 GHz and often even below 1 GHz, where a large legacy of deployed cellular communication systems will remain for a long time to come. Therefore, an important goal will be a "spectrum-compatible" interface that provides best coexistence with legacy radio technologies.

Overall Architecture

Figure 5 illustrates the high-level 5G architecture containing the 5G radio-access functionality supporting the evolution of LTE, as well as the higher frequency 5G radio access technologies. Also illustrated is the 5G core network functionality supporting 5G access while integrating the evolution of legacy access (e.g., 2G, 3G), as well as fixed access. The architecture is supported by a common 5G network management and transport functionality.

The key architecture challenge for the 5G radio access architecture is to integrate the different access technologies and provide effortless and seamless mobility for the end user when transitioning between technologies. The architecture should provide the operator with a single integrated network that achieves high resource efficiency by pooling radio and network resources, and attains high end-user performance by using access aggregation when applicable. The one-network approach is also important for efficient operation and management, reducing operating expenses and providing simple migration paths for increasing network performance. Examples that enable efficient integration of the 5G access technologies include multi-connectivity approaches where the terminal is simultaneously connected on several 5G access technologies or frequency bands. This makes simultaneous data transmission and reception possible on multiple layers or, alternatively, quick failover in case the connection to one layer is lost.

The key challenge related to the 5G core-network architecture is to make it possible to address new 5G use cases, such as mission-critical MTC and ultra-low-latency applications, currently not addressed by cellular networks. In addition, there would be support for optimizing existing use cases, such as media distribution, indoor networks, massive MTC and so on. It is foreseen that supporting these different use cases will lead to an increased need for flexibility in how the network functions and service layer enablers are deployed and operated.

For example, some scenarios might require core network- and service-layer functions to be deployed closer to the radio access to provide excellent end-to-end latency performance and support local communications between users at the same site. In other scenarios some parts of the network can be shared with other operators or with an enterprise or site owner, but shared parts of the network must still be integrated with the rest of the operator network flexibly and seamlessly. This also puts strict requirements on having good security solutions supporting the separation of networks and users in different security domains. It is expected that the 5G core network will utilize the ongoing evolution of software defined networks (SDNs) and network

function virtualization (NFV) to provide a high level of flexibility and scalability when supporting 5G deployments.

The evolved 5G network should also provide service enablers and optimizations yielding added benefits for network integrated services compared to pure over-the-top services. These service enablers could, for example, include mechanisms for reducing battery consumption for MTC devices by enabling longer sleep cycles and connectivity procedures with reduced overhead, or providing a higher degree of reliability for mission-critical MTC. Other service enablers or optimizations could include support for more efficient media distribution.

Summary

5G wireless systems will enable the diverse communication needs of the Networked Society, providing access to information and sharing of data anywhere and anytime for anyone and anything. This will be achieved by a combination of the evolution of existing wireless systems, especially LTE, and complementary new radio-access technologies operating at higher frequencies.

References

[1] J. Thompson *et al.*, "5G Wireless Communication Systems: Prospects and Challenges," *IEEE Commun. Mag.*, Feb. 2014, pp. 62–64.

[2] METIS, "Mobile and Wireless Communication Enablers for the Twenty-Twenty Information Society," Feb. 2013, https://www.metis2020.com/wp-content/uploads/2012/10/METIS_factSheet_2013.pdf.

[3] R. Baldemair *et al.*, "Evolving Wireless Communications: Addressing the Challenges and Expectations of the Future," *IEEE Vehic. Tech. Mag.*, vol. 8, no. 1, pp. 24–30, Mar. 2013.

[4] M. Fallgren (ed.) *et al.*, "Scenarios, Requirements and KPIs for 5G Mobile and Wireless System," METIS Deliverable D1.1, Apr. 2013, https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D1.1_v1.pdf.

[5] ITU-R Recommendation M.1457–11, "Detailed Specifications of the Radio Interfaces of International Mobile Telecommunications-2000 (IMT-2000)," Feb. 2013.

[6] ITU-R Recommendation ITU-R M.2012. "Detailed Specifications of the Terrestrial Radio Interfaces of International Mobile Telecommunications Advanced (IMT-Advanced)," Jan. 2012.

[7] ITU-R WP5D, Working Document Toward Preliminary Draft New Recommendation ITU-R M.[IMT.VISION], "IMT Vision – "Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Document 5D/615 Attachment 3.8.

[8] ITU-R WP5D, Working Document Towards A Preliminary Draft New Report ITU-R M.[IMT.FUTURE TECHNOLOGY TRENDS], "Future Technology Trends of Terrestrial IMT Systems," Document 5D/615 Attachment 5.2.

[9] ITU-R WP5D, Working Document Towards A Preliminary Draft New Report ITU-R M.[IMT.ABOVE 6 GHz], "The Technical Feasibility of IMT in the Bands Above 6 GHz," Document 5D/615 Attachment 5.10.

[10] T. Tirronen *et al.*, "Machine-to-Machine Communication with Long-term Evolution With Reduced Device Energy Consumption," *Trans. Emerging Telecommun. Technologies*, vol. 24, no. 4, June 2013, pp. 413–26.

Biographies

Erik Dahlman is a senior expert in radio-access technologies within Ericsson Research, Ericsson AB, in Stockholm, Sweden. He was deeply involved in the development and standardization of 3G wireless access. Later he was involved in the standardization/development of 4G (LTE) wireless access and its continued evolution. He currently focuses on research and development of future 5G wireless access. He is the co-author of the book, "3G Evolution–HSPA and LTE for Mobile Broadband," and its follow-up, "4G-LTE and LTE-Advanced for Mobile Broadband." He is frequent speaker at international conferences and holds more than 100 patents in the area of mobile communications. In 2009 he received the Swedish Government Major Technical Award for contributions to the technical and commercial success of HSPA. In the spring of 2014 he and colleagues at Ericsson were nominated for the European Inventor Award, the most prestigious inventor award in Europe, for their contributions to LTE.

Gunnar Mildh received his M.Sc. in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2000. In the same year, he joined Ericsson Research, Ericsson AB, Stockholm, and has since been working on standardization and concept development for GSM/EDGE, HSPA and LTE. His focus is on radio network architecture and protocols. He is currently an expert in radio network architecture at the Wireless Access Networks Department of Ericsson Research.

Stefan Parkvall [SM] is currently a principal researcher at Ericsson Research, researching future approaches to radio access. He is one of the key individuals in the development of HSPA, LTE and LTE-Advanced, served as an IEEE Distinguished lecturer 2011-2012, and is co-author of the popular books, "3G Evolution–HSPA and LTE for Mobile Broadband" and "4G–LTE/LTE-Advanced for Mobile Broadband." In 2009, he received the Swedish government Major Technical Award for his work on HSPA, and in 2014 he and colleagues at Ericsson were nominated for the European Inventor Award, the most prestigious inventor award in Europe, for their contributions to LTE. Dr Parkvall received the Ph.D. degree in electrical engineering from the Royal Institute of Technology in 1996. His previous positions include assistant professor in communication theory at the Royal Institute of Technology, Stockholm, Sweden, and a visiting researcher at University of California, San Diego, USA.

Janne Peisa has been working at Ericsson Research on the research and development of 3G, 4G and 5G systems since 1998. Previously, he coordinated Ericsson's radio-access network standardization activities in 3GPP, and currently he leads the Ericsson Research 5G systems program. He has authored several publications and patents and holds both an M.Sc. and a Ph.D. from the University of Helsinki, Finland.

Joachim Sachs is a principal researcher at Ericsson Research working on future wireless communication systems. After studies in Germany, France, Norway and Scotland, he received diploma and doctorate degrees from Aachen University and the Technical University of Berlin, Germany, respectively. In 2009 he was a visiting scholar at Stanford University. Since 1995 Joachim has been active in the IEEE and the German VDE Information Technology Society (ITG), where he currently co-chairs the technical committee on communication networks and systems.

Yngve Selén (yngve.selen@ericsson.com) joined Ericsson Research in 2007 after completing his Ph.D. in signal processing at Uppsala University in Sweden the same year. He currently holds a master researcher position at Ericsson and has been involved in future radio access and 5G research for several years, both as an active researcher and as a project manager.

Johan Sköld [SM] is currently a principal researcher at Ericsson Research and has been working on the evolution and standardization of 2G, 3G, 4G and 5G mobile systems since 1989, mainly in the areas of RF requirements and system performance. He is co-author of the popular "3G Evolution–HSPA and LTE for Mobile Broadband" and "4G–LTE/LTE-Advanced for Mobile Broadband." Sköld holds M.Sc. degrees in electrical engineering from the Royal Institute of Technology, Stockholm, and the University of Washington, Seattle.

It is expected that the 5G core network will utilize the ongoing evolution of software defined networks and network functions virtualization to provide a high level of flexibility and scalability when supporting 5G deployments.

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 stake holders 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 of 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 it. This 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, 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 lifecycle

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.