IEEE COMMUNICATIONS MAGAZINE

April 2017, Vol. 55, No. 4

- SDN Use Cases for Service Provider Networks
- Integrated Circuits for Communications
- Fog Computing and Networking
- Design and Implementation

IEEE ComSoc IEEE Communications Society

A Publication of the IEEE Communications Society

www.comsoc.org

IEEE COMMUNICATIONS MAGAZINE

APRIL 2017, vol. 55, no. 4
www.comsoc.org/commag

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: $71: print, digital, and electronic. $33: digital and electronic. $1001: non-member print.

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

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

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

ADVERTISING: Advertising is accepted at the discretion of the publisher. Address correspondence to: Advertising Manager, *IEEE Communications Magazine*, IEEE, 445 Hoes Lane, Piscataway, NJ 08855-1331.

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 submitted 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. All submissions will be peer reviewed. For further information contact Tarek El-Bawab, Associate Editor-in-Chief (telbawab@ieee.org).

BPA WORLDWIDE

CURRENTLY SCHEDULED TOPICS

Topic	Issue Date	Manuscript Due Date
Emerging Trends, Issues and Challenges in Big Data and Its Implementation towards Future Smart Cities	December 2017	April 20, 2017
Human-Driven Edge Computing and Communication	November 2017	April 30, 2017
Amateur Drone Surveillance: Applications, Architectures, Enabling Technologies, and Public Safety Issues	January 2018	May 1, 2017
Education & Training: Scholarship of Teaching and Supervision	November 2017	May 1, 2017
Heterogeneous Ultra Dense Networks	December 2017	May 15, 2017
Imminent Communication Technologies for Smart Communities	January 2018	June 1, 2017
Mobile Big Data for Urban Analytics	January 2018	June 1, 2017
Point-to-Multipoint Communications and Broadcasting in 5G	March 2018	June 30, 2017
Advanced Industrial Wireless Sensor Networks and Intelligent IoT	February 2018	July 1, 2017
Multi-Channel Cognitive Radio Ad Hoc Networks	April 2018	August 1, 2017

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

IEEE ComSoc™
IEEE Communications Society

Join our Community!

IEEE

Networking • Conference Discounts • Technical Publications • Volunteer

Special Member Rates

50% off Membership for new members.

Offer valid March through 15 August 2017.

Member Benefits and Discounts

Valuable discounts on IEEE ComSoc conferences

ComSoc members save on average $200 on ComSoc-sponsored conferences.

Free subscriptions to highly ranked publications*

You'll get digital access to IEEE Communications Magazine, IEEE Communications Surveys and Tutorials, IEEE Journal of Lightwave Technology, IEEE/OSA Journal of Optical Communications and Networking and may other publications – every month!

*2015 Journal Citation Reports (JCR)

IEEE WCET Certification program

Grow your career and gain valuable knowledge by Completing this certification program. ComSoc members save $100.

IEEE ComSoc Training courses

Learn from industry experts and earn IEEE Continuing Education Units (CEUs) / Professional Development Hours (PDHs). ComSoc members can save over $80.

Exclusive Events in Emerging Technologies

Attend events held around the world on 5G, IoT, Fog Computing, SDN and more! ComSoc members can save over $60.

If your technical interests are in communications, we encourage you to join the IEEE Communications Society (IEEE ComSoc) to take advantage of the numerous opportunities available to our members.

Join today at www.comsoc.org

COMMUNICATIONS HISTORY: THE PAST AS A GUIDE TO THE FUTURE

Harvey Freeman

Steve Weinstein

The IEEE Communications Society, or ComSoc, is currently the third largest Society (32,000) within the 425,000 member IEEE. The IEEE was formed on January 1, 1963 from the combination of the American Institute of Electrical Engineers (founded in 1884) and the Institute of Radio Engineers (founded in 1912). From an IEEE Group on Communications Technology, ComSoc was approved for elevation to Society status in the fall of 1971, and officially began operations on January 1, 1972 (65 years ago). As we celebrate our 65th Anniversary this year, it is fitting to revisit our past and learn from it as we move forward in the ever expanding age of communications technology. And the ComSoc person leading this effort is Stephen B. Weinstein (Steve), Chair of our Communications History Committee. Steve received his Ph.D. degree in electrical engineering from the University of California at Berkeley and began a career with Bell Laboratories, American Express, Bellcore (Telcordia), and NEC Research Labs America. Now mostly retired, he lives in New York City and consults part time for patent law firms and the communications industry. He is a Life Fellow of the IEEE, a past ComSoc President (1996–97), and was an early founding Editor-in-Chief of *IEEE Communications Magazine*. He received the IEEE's 2016 Richard M.Emberson Award for "contributions to IEEE Publications, Awards, and Globalization." Steve is best known for early research and development on data-driven echo cancellation and generation of OFDM signals using the fast Fourier transform. He received the 2006 Eduard Rhein Foundation (Germany) basic research prize for his fundamental OFDM work.

COMSOC'S COMMUNICATIONS HISTORY ACTIVITIES

The past 150 years of discovery, invention, and deployment of communications technology have profoundly changed society and enhanced the quality of life. Electronic communication made possible the web of information exchanges that now define our lives. The history of electronic communication, including the history of our own IEEE Communications Society, is not only a fascinating story but also a guide to the social and economic processes by which ideas become research, research and commercial interests stimulate development, and development, when the time is right, leads to wide use of devices and services. Knowing this past enriches our understanding of how we came to where we are, and this understanding helps uncover opportunities for each of us to influence the future.

The IEEE has a history committee (https://www.ieee.org/about/history_center/history_committee.html) with a paid staff that operates a major electrical engineering history center. In addition to maintaining extensive archives on both technologies and the history of the IEEE itself, including recorded oral histories from noted IEEE members, the IEEE History Center sponsors celebrations of historic "milestones" across IEEE's fields of interest. A book currently in preparation describes the origins of the IEEE in its two predecessor societies, the older AIEE and the younger IRE, that merged in 1963, as noted above.

ComSoc is one of several IEEE societies that sponsor their own, more specialized history activities, in our case the ComSoc Communications History Committee currently chaired by me. I originally joined the IRE and, having reached the age when I might be considered historical, work with a very few additional Committee members to generate history articles for *IEEE Communications Magazine* and *IEEE Wireless Communications Magazine*, sponsor history sessions at major ComSoc conferences, and produce other materials on appropriate occasions. The most recent article and session were, respectively, "History of Radio Propagation" by Jorgen Andersen in the February 2017 issue of *Communications Magazine*, and the "History of Sensor Networks" panel session at Globecom 2016.

A substantial number of past History articles, most published in *Communications Magazine*, are available on the Communication History Committee's web page (comsoc.org/about/communications-history). My predecessor as Chair of the Communications History Committee, Mischa Schwartz, published two on "Improving the Noise Performance of Communication Systems," recalling first the initial breakthroughs of the 1920s, and then the major advances of the 1930s and early 1940s. The prolific Mischa also wrote "Carrier-Wave Telephony over Power Lines: Early History," "Armstrong's Invention of Noise-Suppressing FM," a story of technical brilliance and legal conflicts, and "The Origins of Carrier Multiplexing: Major George Owen Squier and AT&T."

Jerry Hayes conveyed the romance of laying undersea cable on stormy seas in "A History of Transatlantic Cables." Joel Engel brought back "The Early History of Cellular Telephony," recalling the AMPS days. Leonard Kleinrock wrote "An Early History of the Internet" in which he played a significant part. Fred Andrews described "Early T-Carrier History," the introduction of digital access into the telephone network before everyone understood the great advantages of digital communication, and John Cioffi published "Lighting Up Copper," about very high speed Digital Subscriber Line. Dave Falconer provided an authoritative overview of critical improvements in data communication in his "History of Equalization 1860-1980." Norm Abramson contributed "The AlohaNet-Surfing for Wireless Data," telling the story of that very innovative experiment that preceded the Ethernet and numerous wireless services. Hans Peek wrote "The Emergence of the Compact Disc," explaining the international cooperation and advanced optical recording techniques that went into that successful consumer product. The significant role of a Russian scientist in

early radio communication, not so well known in most of the world because his Russian Navy client classified the work, was described by Orest Vendik in "Significant Contribution to the Development of Wireless Communication by Professor Alexander Popov." There have been other history papers as well describing significant advances made in various regions of the world, and in particular in Europe's GSM cellular mobile system, TRANSPAC in France, and computer networking in Korea.

I wrote "The History of Orthogonal Frequency-Division Multiplexing," noting the early deployments in military VHF radios in the 1960s, and even the contribution to the Fast Fourier Transform made by Carl Friedrich Gauss in 1805. Several other articles are in preparation, including "History of SDN," "History of Echo Cancellation," "History of Deep Space Communication," "History of MIMO," and "History of Coding," all by prominent members of our community. Readers of this column are encouraged to submit proposals for topics and authors, not excluding themselves.

The most significant examples of "special occasion" materials were *A Brief History of Communications*, a concise book that had two editions, the first in 2002 for ComSoc's 50th anniversary and the second in 2012 for our 60th anniversary, and a video described below. The book, available for free downloading on the Communications History Committee's web page, reviews communications technology from "the beginning" up to 2012. It then describes the development of the IEEE Communications Society from its founding as the IRE Professional Group on Communications Systems in 1952 through its six decades up to 2012. It further includes transcripts of interviews, "oral histories" made by the IEEE History Center with many prominent contributors to communications technologies and to ComSoc.

A 23-minute video, released in 2012, of past ComSoc Presidents offering very brief reminiscences about the important events they lived through during their professional and ComSoc careers is also available on the web page. It is fascinating to hear these first-person accounts illuminating past trends, breakthroughs and failures, and the feelings of our Past Presidents about membership in a global communications engineering community.

The Communications History Committee is an example of many ComSoc volunteer activities that cost next to nothing but add meaningfully to our shared cultural heritage. Busy with our obligations and careers, it can be difficult to find the time and energy to write an article about the past, but these articles are a great service to our fellow ComSoc members. We encourage you to explore the history of our fields of interest and the Communications Society itself.

SOCIETY MEMBERS NAMED TO IEEE FELLOW GRADE

Election to the grade of IEEE Fellow is one of the highest honors that can be bestowed upon our members by the Institute in recognition of their technical, educational, and leadership achievements. Only a select few IEEE members earn this prestigious honor.

Congratulations to the following Communications Society members for their election to the grade of Fellow of the IEEE. They now join company with a truly distinguished roster of colleagues.

RAVIRAJ ADVE

For development of signal processing techniques for airborne radar.

N. ASOKAN

For contributions to system security and privacy.

MICHAEL BUEHRER

For contributions to wideband signal processing in communications and geolocation.

YAN CHEN

For contributions to design, measurement, and security of networking systems.

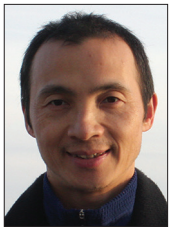
CHRISTOPHER R COLE

For contributions to 10G, 40G, and 100G Optical Ethernet and OTN interfaces.

HUAIYU DAI

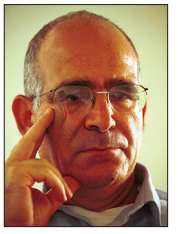
For contributions to MIMO communications and wireless security.

JING DENG



For development and optimization of wireless security and networking protocols.

DANNY DOLEV



For development of consistent, robust, and efficient distributed computing and storage.

FALKO DRESSLER

For contributions to adaptive and self-organizing communication protocols in sensor and vehicular networks.

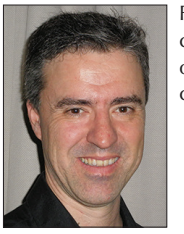
EYLEM EKICI

For contributions to algorithms, protocols, and architectures of multi-hop wireless networks.

MICHAEL GASTPAR

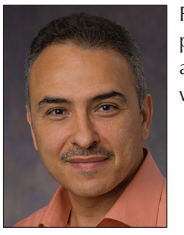
For contributions to network information theory.

STEPHEN HANLY



For contributions to capacity analysis and optimization of wireless communication networks.

HOSSAM HASSANEIN



For contributions to protocols, architectures and analysis of multihop wireless networks.

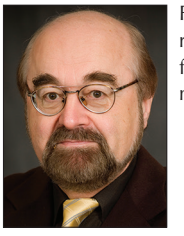
MICHAEL ISNARDI

For contributions to compliance testing and vision-based video compression technologies.

YUCHEUN JOU

For leadership in digital cellular systems and smart mobile devices.

WITOLD KRZYMIEN



For contributions to radio resource management for cellular systems and networks.

QILIAN LIANG

For contributions to interval type-2 fuzzy logic systems.

XIANG LIU

For contributions to broadband optical fiber communication systems and networks.

BORIVOJE NIKOLIC

For contributions to energy-efficient design of digital and mixed-signal circuits.

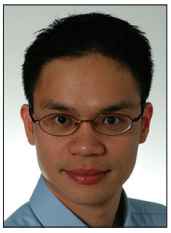
TENG-JOON LIM

For contributions to statistical signal processing in wireless communications.

YONG LIU

For contributions to multimedia networking.

DUSIT NIYATO



For contributions to resource allocation in cognitive radio and cellular wireless networks.

PHONE LIN

For contributions to resource management and service development for mobile networks.

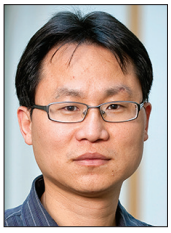
JOSEPH BRYAN LYLES

For contributions in local network technology, Internet measurement, and research cyberinfrastructure.

JUNG-MIN PARK

For contributions to dynamic spectrum sharing, cognitive radio networks, and security issues.

XIAODONG LIN

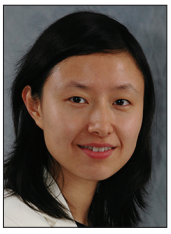


For contributions in secure and privacy-preserving vehicular communications.

ARUMUGAM NALLANATHAN

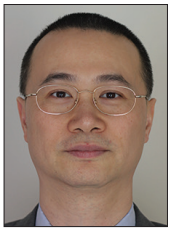
For contributions to cooperative communications and cognitive radio networks.

LILI QIU



For contributions to wireless network management.

XIAOJUN LIN



For contributions to scheduling and control of wireless networks.

SHU NAMIKI

For contributions to optical amplification.

REZA REJAIE

For contributions to multimedia and peer-to-peer networking, and Internet measurement.

JIANGCHUAN LIU

For contributions to multimedia communications and content distribution over the Internet.

PANOS NASIOPOULOS

For leadership in DVD authoring and digital multimedia technologies.

MICHAEL RICE

For contributions to communication waveforms, detection algorithms, and channel models for aeronautical telemetry.

SEB SAVORY

For contributions to digital coherent transceivers for optical fiber communication.

BEHZAD SHAHRARAY

For leadership in content-based processing and retrieval of multimedia information.

HAMID SHARIF

For development of railroad wireless communication.

PRASUN SINHA

For contributions to scheduling and resource allocation in wireless networks.

FREDRIK TUFVESSON

For contributions to measurement and modeling of wireless propagation channels.

JAAP VANDEBEEK

For contributions to orthogonal frequency division multiplexing.

CHENG-XIANG WANG

For contributions to wireless channel modeling for vehicular networks.

CHONGGANG WANG

For contributions to Internet of Things enabling technologies.

JIANGZHOU WANG

For contributions to multiple access and resource allocation in wireless mobile communications.

WENYE WANG

For contributions to modeling and performance evaluation of wireless networks.

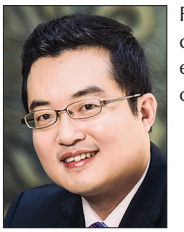
XIANBIN WANG

For contributions to OFDM systems and distributed transmission technologies.

HALIM YANIKOMEROGLU

For contributions to wireless access architectures in cellular networks.

RUI ZHANG



For contributions to cognitive radio and energy harvesting communications.

CE ZHU

For contributions to video coding and communications.

PHOTOS NOT AVAILABLE

ISRAEL CIDON

For contributions to high-speed packet networks, network-on-chip and wide area file systems.

PABLO ESTEVEZ

For contributions to feature selection and visualization of large data sets.

AMIR KHANDANI

For contributions to resource allocation and interference management in network information theory.

IEEE ICC 2017 ANNOUNCES ITS INDUSTRY PROGRAM

HUNDREDS OF COMMUNICATIONS EXPERTS TO DISCUSS NEXT WAVE 5G, IOT, AI, BIG DATA AND SPECTRUM SHARING INNOVATIONS IN PARIS, FRANCE FROM 21-25 MAY

BY JAMSHID KHUN-JUSH (QUALCOMM) AND LUIS M CORREIA (IST–UNIVERSITY OF LISBON)
IEEE ICC 2017 INDUSTRY FORUMS AND EXHIBITIONS CO-CHAIRS

IEEE ICC 2017 (http://icc2017.ieeeicc.org), the leading international venue dedicated to the advancement of wireless and wireline communications worldwide, will host nearly 20 industry sessions devoted to the latest developments in next wave technologies such as 5G, IoT, AI, Big Data and Spectrum Sharing. Held this year from 21-25 May at the Paris Congress Center (Palais des Congrès) in Paris, France, IEEE ICC is recognized globally as an annual showcase of innovation exploring the latest in communications technologies and policies.

This year, the Industry Program will address a wide array of topics ranging from standardization to propagation challenges, including security, artificial intelligence in networks, and spectrum, among others, most of them addressing 5G, but not entirely. The program will begin on Monday, 22 May with two panels on 5G and IoT, followed by the CTO Forum addressing "From Myth to Reality: Rapid Deployment at Scale Toward 5G Global Success." This forum brings together high-level representatives from leading mobile operators and manufacturers to critically assess such measures, from the pace of standardization to over-the-air trials and interoperability testing for validation, in order to drive the ecosystem toward rapid and wide-scale commercialization, which are the basis for making 5G a success. The session is intended to be an open and frank debate among developers, operators and vendors about the timeline and potentials for 5G to address the larger vision beyond 2020.

These highly focused panels will continue through Wednesday, 24 May with a broad range of experts from operators, vendors, service providers, and consultants, among other industry players, as well as academia, research centers, standardization bodies and regulators. Discussions center on topics such as:

- Big Data Analytics for Smart and Connected Health
- 5G C-RAN and X-Haul
- Spectrum Sharing in Future Wireless Communications Systems
- ComTech Innovation and Trends Toward Smart Connected Car Environment
- Challenges and Opportunities in SDN/NFV and 5G Security
- 5G: What Architecture to Serve Vertical Industries?
- The Path to and Obstacles in Network Virtualization
- Propagation Challenges in New 5G Use Cases
- 5G Progress and Challenges
- Network AI

Attendees engaged at a previous conference.

The program also offers two seminars dedicated to simulation tools, and measurement equipment and testbeds. One, "The Road to 5G: Design, Simulate and Prototype Wireless Systems using MATLAB and Simulink," will be given by John Wang of MathWorks. The talk will discuss how MATLAB and Simulink provide an integrated environment for designing, simulating, and prototyping 5G wireless systems, covering up-to-date progress on MATLAB and Simulink in the 5G area and how it can address engineers' and researchers' challenges in 5G areas, such as mm Waves, Massive MIMO, and prototyping. The other, "Real-Time Prototyping of Massive MIMO," will be presented by Douglas Kim of National Instruments. It will provide insights into Massive MIMO and rarely shared information on the implementation of such a complex system, presenting information on the practical challenges associated with the implementation of such systems, and a close examination of a test bed, using real-world over-the-air channels and true real-time data rates.

Other significant events to be held during the heart of the conference will include the presentation of nearly 1,200 original scientific papers as well as the keynotes of leading industry executives and researchers. For instance, during the opening and keynote session, Marcus Weldon, President of Bell Labs and CTO of Nokia, will share insights on driving technological and architectural innovations. Over next few days Marcus will be joined by:

- Georges Karam, Founder, Chairman and CEO of Sequans Communications, who will speak on "4G Wireless Enabling the Internet of Things (IoT)."
- Wen Tong, Huawei Fellow, Head of Huawei Wireless Research and CTO of Huawei Wireless, who will detail "The Blueprint of 5G–A Global Standard."
- Alain Aspect, Director of Research at CNRS and Professor at Ecole Polytechnique, who will talk about "From Einstein's Doubts to Quantum Information: A New Quantum Revolution."
- Serge Willenegger, Senior Vice President of Qualcomm Technologies, who will address "The On the Path to 5G."
- Giuseppe Caire, Professor at Technical University of Berlin, who will discuss "Wireless Edge Caching: Promises and Recent Advances."

For more information about IEEE ICC 2017 including program updates and registration information, please visit http://icc2017.ieee-icc.org. All website visitors are also invited to network with colleagues and peers, share their professional experiences through the conference's Facebook, LinkedIn and Twitter pages.

ICIN 2017: INNOVATIONS IN CLOUDS, INTERNET, AND NETWORKS

ANTONIO MANZALINI, TELECOM ITALIA MOBILE
NOEL CRESPI, INSTITUT-MINES TÉLÉCOM, TÉLÉCOM SUDPARIS EVRY, FRANCE

The 20th ICIN conference took place in Paris, France, on March 7-9, with the technical co-sponsorship of the IEEE and IEEE Communications Society, and the support of Nokia, UPMC, Gandi, and Institut Mines-Telecom. The conference was attended by 110 delegates from 21 countries and representing 55 different organizations. Attendees participated to discuss innovative technologies and solutions for clouds, Internet, and networks. ICIN2017 has a 28-year history of anticipating key trends and innovation avenues in telecommunications that are essential for developing and provisioning ICT services.

ICIN 2017 embraced several main aspects of the ongoing digital business transformation of telecommunications, called "softwarization," which is steering the evolution of both networks and service platforms (e.g., through cloud and edge computing architectures) even up to future terminals, machines, and smart objects (e.g., through fog computing). ICIN 2017 focused on some key areas of this systemic transformation: architectural delayering, simplifications and optimization of the operational processes (e.g., through orchestrators and new paradigms of OSS/BSS capable of managing the complexity and heterogeneity of SDN and NFV infrastructures), development of flexible and programmable "platforms of platforms," and the exploitation of big data by means of data mining and cognitive technologies. Mastering the software was recognized as a must for a successful digital business transformation. In addition, standardization and open source software solutions and interfaces were still seen to be hot issues.

The paper acceptance rate of the conference was 32%. The 31 accepted papers were organized into four technical tracks: Network and Service IT-zation (chaired by Galis from UCL, UK); Internet of Things (chaired by Luigi Atzori from the University of Cagliari, Italy); Actionable Big Data and Artificial Intelligence (chaired by Albert Cabellos, from UPC, Spain); and Control Orchestration and Management and Policy (chaired by Bruno Chatras from Orange Labs, France). The conference included a demonstration and poster tracks, chaired by Roberto Bruschi (CNIT, Italy). There were also two workshops: "Infrastructures and Data Analytics for Smart Cities" (chaired by Ralf Tönjes from the University of Applied Sciences, Osnabrück, Germany, and Edith C.-H. Ngai from Uppsala University, Sweden); and "5G for Universal Access: Meeting the challenges for Urban and Rural Coverage" (chaired by Philip Kelley from Nokia Bell Labs France).

There was also a special session on "Trust and Access Control for the Internet of Things," chaired by Emmanuel Bertin from Orange Labs (France), which highlighted the need for innovative access control models for the IoT.

After a tutorial on ONOS, presented by Andrea Campanella (ON.Lab-ONF), the conference was opened by Antonio Manzalini (TIM, Italy), Chair of the ICIN 2016 Technical Program Committee, and Noël Crespi, General Chair of the Conference and Chair of the ICIN International Advisory Board.

Five keynote speeches were delivered around the theme of the conference.

Dr. Chih-Lin I, Chief Scientist of Wireless Technologies of China Mobile (China), gave the first keynote speech on the topic "The Perfect Storm: IT+CT+DT," emphasizing how the intertwining of the advances in information technology (IT), communication technology (CT), and data technology (DT) will lead to a profound transformation of telecommunications, which will see its first concrete expressions with the 5G. The second keynote speech was from Stephen Terrill (Ericsson, Spain), with the main topic "Multi-Domain Orchestration and Automation in the Age of SDN and NFV." The second day was opened by the keynote speech of Marina Thottan, Director at Nokia Bell Labs (USA). The title of the talk was "Programmable Network Operating System: Creating the Network Brain." During her speech, Thottan first addressed the requirements of a network operating system for future "softwarized" telecommunications infrastructures, then presented the NetGraph data model applied to multilayer carrier networks and eventually the prototype of the Network Brain being developed at Nokia Bell Labs. The third day was opened with the keynote speech by Marie-Paule Odini, Director in HPE (France), and Distinguished Technologist in the Communication Solution Business. The talk addressed the status, challenges, and open issues related to NFV evolution toward 5G. The fifth keynote speech was by Thomas Michael Bohnert, Professor at Zurich University of Applied Sciences (Switzerland), who gave a talk on cloud robotics.

Also this year, ICIN2017 consistently previewed trends in networks and services. Keynotes and papers highlighted the significant innovation coming from the "softwarization" of telecommunications, a systemic transformation based on the convergence of enabling technologies such as SDN and NFV, but also the evolution of cloud computing toward edge and fog computing and the revamping of artificial intelligence.

Planning for ICIN 2018 is already underway under the leadership of Prosper Chemouil (General Chair, Orange), Laurent Ciavaglia (TPC Chair, Nokia), Rahim Tafazolli (TPC Chair, University of Surrey), and Noel Crespi (IAB Chair, IMT-TSP). The conference will take place in Paris on February 20-22 2018. For more information please visit http://www.icin-conference.org/

GCN GLOBAL COMMUNICATIONS NEWSLETTER

April 2017
ISSN 2374-1082

MEMBERSHIP SERVICES

North America Region

Interview with T. Scott Atkinson, Director of the NA Region

By Stefano Bregni, Vice-President for Member and Global Activities, and T. Scott Atkinson, Director of the NA Region

This is the fourth article in the series started in November 2016 and published monthly in the IEEE ComSoc Global Communications Newsletter, which covers all areas of IEEE ComSoc Member and Global Activities. In this series of articles, I introduce the six MGA Directors (namely: Sister and Related Societies; Membership Services; AP, NA, LA, EMEA Regions) and the two Chairs of the Women in Communications Engineering (WICE) and Young Professionals (YP) Standing Committees. In each article, one by one they present their sector activities and plans.

In this issue, I interview T. Scott Atkinson, Director of the North America Region (NA).

Stefano Bregni

T. Scott Atkinson

ComSoc Membership		
Region	Members	Percent of total
NA	12,211	42.2
EA	6,787	23.5
LA	1,149	4.0
AP	8,788	30.4
Total	28,935	100.0

Table 1. Shares of ComSoc membership among the four Regions.

Scott is retired and spends most of his time volunteering for the IEEE. He received the B. S. degree in physics, with minors in mathematics and chemistry, from Texas A&I University, Kingsville, TX, USA in 1961. From 1961 to 1967, he was a communications officer in the United States Air Force. In 1967, he worked as a communications engineer with Lockheed Electronics Company on a contract with the NASA Manned Spacecraft Center in Houston, TX. In that position, he supported the testing of the Apollo spacecraft communications sub-system. In 1968, he joined the IEEE and the Communications Society and became a volunteer supporting Section and Chapter activities in Houston, TX. In 1973, he joined Tenneco Inc, Houston, TX, spending 14 years performing engineering tasks on their telecommunication systems. His last major work activity was as a senior technical support specialist for the United Services Automobile Association (USAA) in San Antonio, TX. Retiring in 1997, he now volunteers full time for the IEEE, the IEEE Communications Society, and IEEE Region 5.

I have known Scott since a long time, as one of the most dedicated and best appreciated volunteers in the entire NA Region of the IEEE Communications Society. It is a pleasure for me to interview Scott today and offer him this opportunity to outline his current activities and plans as Director of the NA Region.

Bregni: Scott, would you introduce briefly the North America Region Board?

Atkinson: Our Region consists of the 92 Communications Society Chapters in IEEE Regions 1 through 7, which include all of the United States and Canada.

Per the ComSoc Policies and Procedures, our Board "is responsible for stimulating, coordinating and promoting the activities of ComSoc members and chapters throughout the North America Region." Therefore, we support the activities of the 92 Chapters and its members.

Our NA Board consists of individual IEEE Region Representatives and several other volunteers, namely:

- IEEE Region Representatives Ali Abedi (R1), Kafi Hassan (R2), Scott Midkiff (R3), William Ashe (R4), Fawzi Behmann (R5), Upkar Dhaliwal (R6), and Wahab Almuhtadi (R7).
- NAR Board Vice Chair Fawzi Behmann.
- Distinguished Lecturer/Distinguished Speaker Coordinator Zafar Taqvi.
- Past Chair Merrily Hartmann.
- Advisors Richard Miller, Anader Benyamin-Seeyar, Paul Cotae and John Lyons.

In addition, we have set up an Awards Committee.

Meetings of the Board are held monthly via telecon and twice a year during the ICC and GLOBECOM Conferences. In short, what are the activities of the NA Region which the NAR Board oversees?

Bregni: What is your special focus?

Atkinson: The Board oversees the ComSoc activities at the Chapter level through each of the Region Representatives, maintaining contact with each of their IEEE Region Chapters to assist them by suggesting activities and resolving any issues. Special focus is given when a chapter has not shown any recent activity as shown on the IEEE L31 reports. Additionally, at the direction of the Director, each Regional

(Continued on Newsletter page 4)

Region	Higher grade	Percent of total	Senior grade	Percent of total	Member grade	Percent of total	Student grade	Percent of total
North America	11.164	44.1%	2.184	47.7%	7.214	39.3%	873	27.2%
European Area	5.813	23.0%	1.178	25.7%	4.333	23.6%	881	27.4%
Latin America	845	3.3%	134	2.9%	693	3.8%	291	9.1%
Asia Pacific	7.505	29.6%	1.083	23.7%	6.108	33.3%	1169	36.4%
Total	25.327	100%	4.3579	100%	18.348	100%	3214	100%

Table 2. Members by category in the four Regions.

CONFERENCE REPORT

2016 International Conference on Advanced Technologies for Communications (ATC2016), Hanoi, Vietnam

Organized Jointly by the IEEE ComSoc Sister Society Radio Electronics Association of Vietnam (REV) and the IEEE Communications Society

By Duc-Tan Tran, VNU-UET, Viet Nam

The 2016 International Conference on Advanced Technologies for Communications (ATC2016) was jointly organized by The IEEE Communications Society, The Radio-Electronics Association of Vietnam (REV), and VNU University of Engineering and Technology in Hanoi, Vietnam, on October 12-14, 2016. The technical program of the conference featured 86 papers, including 81 oral presentations and five poster presentations. All papers are indexed by IEEEXplore.

The IEEE Communications Society was actively involved in the conference organizing committee with Prof. Stefano Bregni (IEEE ComSoc Vice-President for Member and Global Activities) acting as the Technical Program Co-Chair, Prof. Vijay Bhargava (2012–2013 President of IEEE ComSoc) as the Steering Co-chair, and Prof. Hikmet Sari (2014–2015 IEEE ComSoc Vice-President for Conferences) as the General Co-Chair. The conference was also honored to welcome Prof. José Roberto Boisson de Marca, 2014 President of IEEE, as the Honorary Co-Chair.

ATC2016 attracted more than 200 scientists and researchers from 29 different countries around the world to discuss the latest technologies in electronics and communications. Typical dicussion topics included antennas and propagation, microwave engineering, communications, signal processing, biomedical engineering, networks, IC and electronics design.

Conferring the best paper award to Dr. Pham Ngoc Nam.

The conference was successful in inviting the world's leading researchers to deliver three important keynote speeches in the area of wireless communications and multimedia. The first speech, entitled "Power Efficiency in Wireless Communications–A Historical Review," was delivered by Prof. Hikmet Sari, Head of the Telecommunications Department at SUPELEC, France, and Chief Scientist of Sequans Communications, France. The next keynote speech was on "Wireless-Powered Communication Networks: Architectures, Protocols, and Applications," given by Prof. Dong In Kim, Director of the Cooperative Wireless Communications Research Center, Sungkyunkwan University (SKKU), Korea. The two speeches garnered the special attention of more than 200 delegates. The third speech was presented by Prof. Minh N. Do, IEEE Fellow, Professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign (UIUC) on the hot topics of "Quantifying and Extracting Visual Information from Mobile Devices." The technical sessions of the conference, with 86 papers, were divided into nine tracks and three special sessions running over three consecutive days. The conference received 174 submissions, of which 86 papers were accepted, for an acceptance rate of 49%. Two excellent papers were selected by the technical committee to receive the best paper awards.

Prof. Hikmet Sari delivered keynote speech.

A technical session.

The ATC2016 conference was a great opportunity for long-term and sustainable collaborations between the scientists in Viet Nam and other countries. The next conference is scheduled for 2017 in Quy Nhon, Viet Nam.

For further information, please see the conference web site http://rev-conf.org/. More photos of ATC2016 are available at http://rev-conf.org/conferences-photos

Opening ceremony.

CHAPTER REPORT

IEEE ComSoc North-America Regional Chapter Chair Congress at IEEE GLOBECOM 2016, Washington, DC, USA

By T. Scott Atkinson, NA Region Director, IEEE ComSoc

We are pleased to present the activities of the recent North America Regional Chapter Chair's Congress held in Washington, DC USA on December 3rd and 4th, 2016

The meeting was approved by our Vice President, Stefano Bregni, and the ComSoc Board of Governors during meetings at the ICC Conference held in Kuala Lumpar, Malaysia, 23-27 May 2016. Immediately thereafter, the North America Region Board began planning our upcoming meeting at IEEE GLOBECOM 2016.

I would first like to introduce our NA Board, where you can find more information on our web site: http://na.regions.comsoc.org/board/. Our Board is composed of the following volunteers:
Chair (NAR Director): T. Scott Atkinson
Vice Chair: Fawzi Behmann
Past Chair: Merrily Hartmann
DLT/DSP/MDSG Coordinator: Zafar Taqvi
IEEE Region 1 Representative: Ali Abedi
IEEE Region 2 Representative: Kafi Hassan
IEEE Region 3 Representative: Scott Midkiff
IEEE Region 4 Representative: William "Bill" Ashe
IEEE Region 5 Representative: Fawzi Behmann
IEEE Region 6 Representative: Upkar Dhaliwal
IEEE Region 7 Representative & NAR Information Technology Coordinator: Wahab Almuhtidi
Advisor (R7 DLT Support): Anadeer Benyamin-Seeyar
Advisor (Conferences): Richard Miller
Advisor: John Lyons
Advisor: Paul Cotae

Our Region is unique in that we cover seven IEEE Regions, whereas each of the other ComSoc Regions have one each, Regions 8, 9 and 10. Thus, we need to have a rather large Board to oversee the activities of the 92 chapters (42% of all ComSoc) in the seven Regions.

We operate on a set of ComSoc Policies and Procedures that have been approved by the ComSoc Member Global Activities Committee and ultimately the ComSoc Board of Governors.

Leading up to the NARCCC at GLOBECOM, we began by holding monthly conference calls to develop an agenda for the meeting and our approach to inviting the chapter to participate. Since we have a large contingent of the ComSoc chapters, we divide them into two areas, the Eastern Area and the Western Area, and focus our NARCCC on each separately. The last NARCCC was held in Austin at the IEEE GLOBECOM 2014 and focused on the Western Area Chapters. So in 2016, our focus was on the Eastern Area Chapters. Additionally, we included a few chapters from Region 5 plus one chapter (Houston) in the process of being reactivated. Overall we had representatives from 24 chapters: Maine (R1)–Ali Abedi; Montreal (R7)–Reza Soleyman; North Jersey (R1)–Amit Patel; Baltimore (R2)–Anna Romaniuk; Jamaica (R3)–Devon Gayle; NJ Coast (R1)–Newman Wilson; Kingston (R7) – Francois Chan; Toronto (R7)–Eman Hammad; Ottawa (R7)–Wahab Almuhtadi; London (R7)–Hao Lin; New York (R1)–Warner Sharkey; Washington (R2) – Debi Siering; Houston (R5)–Russell Roy; Mohawk Valley (R1)–Brian Spink; New Hampshire (R1)–Mary Bzezenski; Princeton Central Jersey (R1)–Ashutosh Dutta; Worcester County (R1)–Sundar Sundaramurthy; Pittsburgh (R2)–Balaji Palanisamy; Austin (R5)–Fawzi Behmann; San Antonio (R5)–Brian Kelley; Tulsa (R5)–Ali Imran; New Orleans (R5)–Richard Miller; San Diego (R6)–Upkar W; Galveston Bay (R5)–Zafar Taqvi.

Our agenda included:
- ComSoc President's Report: Harvey Freeman, President
- Member & Global Activities Report: Stefano Bregni, VP MGA
- ComSoc Financials (Treasurer): Bob Shapiro, Treasurer
- Executive Director's Report: Susan Brooks, Staff
- Technical Activities & Industry Outreach: Adam Greenberg, Staff
- North America Region Report: T. Scott Atkinson, NA Director. During this presentation, Scott passed out a copy of a Vitality Checklist for successful IEEE ComSoc Chapters. This Checklist can be found on our web site under our RCC tab.
- Chapter Achievement Award (Outstanding Chapter): Wahab Almuhtaki, Chair
- Austin Chapter Report: Fawzi Behmann, Chair
- Membership Services: Zhensheng Zhang, MS Director
- Panel Discussion: Merrily Hartmann
- DLT/DSP Updates: Zafar Taqvi
- ComSoc Seminars (5G, Big Data, Iot, Cybersecurity): Ashutosh Dutta
- Second Panel Discussion: Merrily Hartmann

Also at lunch we had a special speaker, Mr. Michel Jansan, Open Net Ombudsman, U.S. Federal Communications Commission.

All presentations are located on our website:
http://na.regions.comsoc.org/RCC.

Additionally, we announced, thru our IEEE Region Representatives, the opportunity for all chapters to nominate someone in their chapter who had provided exemplary service to their chapter in 2016. Six nominations were received and judged by five Board members not associated with those chapters. We required at least three votes for an individual to be chosen. There were three awards that could be given, which included a plaque and a check for $400.00. There were two winners: Raed Abdullah from the Ottawa Chapter, and Reza Soleymani from the Montreal Chapter. Congratulations to these two outstanding supporters of their chapters.

And finally, one of our own long term Society members was selected to become "Dirctor Emeritus" for Region 5. Ross

Group photo of participants to the IEEE ComSoc North-America Regional Chapter Chair Congress at IEEE GLOBECOM 2016, Washington, DC, USA.

MEMBERSHIP SERVICES/*Continued from page 1*

Representative may be appointed Chair of one of the Board Standing Committees.

Bregni: Is it possible to attend the NA Board meetings and collaborate on some of its activities?

Atkinson: Yes, certainly. The NA Board is open to any IEEE ComSoc member from the NA Region at our Board meetings held during each of the two flagship conferences each year, which are usually held during the Technical Sessions of GLOBECOM/ICC. I hope that as many of our members as possible will get involved and serve on one of our Committees, as the best networking often occurs when you are working toward a common goal.

Bregni: Now, would you give us a brief overview of the membership of the NA Region? How are they distributed?

Atkinson: The number of ComSoc members in the NA region is 12,211, and this represents 42.2% of the 28,935 ComSoc members in total (as of Jan 2017). This is the largest share among all Regions, as shown in Table 1. See Table 2 for more details on ComSoc membership and the various categories.

Bregni: In 2016, the IEEE NA Region Chapter Achievement Award was presented to the Ottawa Communications Society Chapter. Could you tell us about this Chapter and what they did to be awarded?

Atkinson: My pleasure. Their chapter official IEEE name is "Joint Chapter of Communications Society, Broadcast Technology Society and Consumer electronics Society."

The Chapter Chair is Dr. Wahab Almuhtadi, P.Eng. He is a professor and coordinator of the Algonquin College–Carleton University Joint Degree Program "Bachelor of Information Technology–Photonics and Laser Technology–BIT-PLT" in the School of Advanced Technology, at Algonquin College. Dr. Almuhtadi is also the R&D Coordinator at the Algonquin College Faculty of Technology and Trades. Dr. Almuhtadi has more than 20 years of university and college teaching experience at both the undergraduate and graduate levels.

During 2016, the Chapter (1) was heavily involved in communications with local industry, academia, Ottawa Tourism and the City of Ottawa in the preparation for hosting Conferences in Ottawa; (2) co-organized with the student branches and the young professionals many events; (3) held five meetings with Distinguished Lecturers plus five meetings with other noted technology speakers; and (4) the Chapter was also heavily involved in community technical activities.

To make the selection for the Chapter Achievement Award, we examined questionnaires of the chapters, who were reporting their activities for the previous year. The Ottawa Chapter was clearly superior among all those submitted.

Visit to the Ottawa Chapter.

Their award was presented during the Luncheon Awards Ceremony at the recent IEEE GLOBECOM 2016 Conference in Washington, DC. To further recognize the Chapter volunteers, I made a personal visit to Ottawa to congratulate the various Chapter Leaders, who played a significant role in the success of the Chapter.

Bregni: You just had your NA Region Chapter Congress last December at GLOBECOM 2016, in Washington, DC. When do you anticipate the next opportunity for another NA Region Chapter Chairs Congress?

Atkinson: The next opportunity to hold an NA Regional Chapter Chairs Congress could be in 2018, most likely at the IEEE International Conference on Communications (ICC) to be held in Kansas City, MO, May 19-25.

Bregni: Scott, in conclusion, what would you recall as main highlights during your term as Director of the North America Region?

Atkinson: Here are some highlights during my term as Director.

We committed extra time and effort to ensuring that we had a fully staffed and active Board. To maintain communications, I frequently communicated via phone calls and emails with the IEEE Region Representatives to keep them active and informed on current activities.

Additionally, we proposed an RCCC in June and delivered a very successful RCCC meeting at the IEEE GLOBECOM 2016 in Washington, DC with 24 Chapters represented. A highlight of the meeting included a luncheon speaker, Michael Janson, the Open-Internet-Ombudsperson for the U. S. Federal Communications Commission.

To complete the year, we accomplished two additional items of note: a personal visit to the chapter winning our Chapter Achievement Award, Ottawa, Canada, to congratulate the Chapter leaders on their achievements; and receiving six nominations for our Exceptional Service Award and provided, for the first time ever, awards to two outstanding chapter individuals in the Montreal and Ottawa chapters.

CHAPTER REPORT/*Continued from page 3*

Anderson was a ComSoc volunteer since 1977, when he was Secretary of the 1980 National Telecommunications Conference in Houston, TX. He went on to be the Chair of the 1986 National Communications Conference, now named IEEE Globecom, and became the director of the flagship conference committee and a member of the ComSoc Board of Governors as their Treasurer. Select members of ComSoc participated in recognizing him at a special dinner in his honor on January 21, 2017 in Houston, TX.

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 ComSoc IEEE Communications Society
www.comsoc.org/gcn
ISSN 2374-1082

UPDATED ON THE COMMUNICATIONS SOCIETY'S WEB SITE

www.comsoc.org/conferences

2017

APRIL

IEEE ISPLC 2017 — IEEE Int'l. Symposium on Power Line Communications and its Applications, 3–5 Apr.
Madrid, Spain
http://isplc2017.ieee-isplc.org/

WTS 2017 — Wireless Telecommunications Symposium, 26–28 Apr.
Chicago, IL
http://www.cpp.edu/~wtsi/

MAY

IEEE INFOCOM 2017 — IEEE Int'l. Conference on Computer Communications, 1–4 May
Atlanta, GA
http://infocom2017.ieee-infocom.org/

ICT 2017 — Int'l. Conference on Telecommunications, 3–5 May
Limassol, Cyprus
http://ict-2017.org/

IFIP/IEEE IM 2017 — IFIP/IEEE Int'l. Symposium on Integrated Network Management, 8–12 May
Lisbon, Portugal
http://im2017.ieee-im.org/

IEEE EIT 2017 — IEEE Int'l. Conference on Electro Information Technology, 14–17 May
Lincoln, NE
http://engineering.unl.edu/eit2017/

ISNCC 2017 — Int'l. Symposium on Networks, Computers and Communications, 17–19 May
Marrakesh, Morocco
http://www.isncc-conf.org/

IEEE ICC 2017 — 2017 IEEE Int'l. Conference on Communications, 21–25 May
Paris, France
http://icc2017.ieee-icc.org/

JUNE

IEEE BlackSeaCom 2017 — IEEE Int'l. Black Sea Conference on Communications and Networking, 5–9 June
Istanbul, Turkey
http://blackseacom2017.ieee-blackseacom.org/

GIoTS 2017 — Global Internet of Things Summit, 6–9 June
Geneva, Switzerland
http://iot.committees.comsoc.org/global-iot-summit-2017/

IEEE CTW 2017 — IEEE Communciation Theory Workshop, 11–14 June
Natatola Bay, Fiji
http://ctw2017.ieee-ctw.org/

IEEE LANMAN 2017 — IEEE Workshop on Local & Metropolitan Area Networks, 12–15 June
Osaka, Japan
http://lanman2017.ieee-lanman.org/

IEEE SECON 2017 — IEEE Int'l. Conference on Sensing, Communication and Networking, 12–14 June
San Diego, CA
http://secon2017.ieee-secon.org/

EuCNC 2017 — European Conference on Networks and Communications, 12–15 June
Oulu, Finland
http://eucnc.eu/?q=node/156

IEEE/ACM IWQOS 2017 — IEEE/ACM Int'l. Symposium on Quality of Service, 14–16 June
Vilanova i la Geltrú, Spain
http://iwqos2017.ieee-iwqos.org/

IEEE CAMAD 2017 — IEEE Int'l. Workshop on Computer Aided Modeling and Design of Communication Links and Networks, 19–21 June
Lund, Sweden
http://weber.itn.liu.se/~vanan11/CAMAD17/

TMA 2017 — Network Traffic Measurement and Analysis Conference, 21–23 June
Dublin, Ireland
http://tma.ifip.org/

NETGAMES 2017 — Annual Workshop on Network and Systems Support for Games, 22–23 June
Taipei, Taiwan
http://netgames2017.web.nitech.ac.jp/

CLEEN 2017 — Int'l. Workshop on Cloud Technologies and Energy Efficiency in Mobile Communication Networks, 22 June
Turin, Italy
http://www.flex5gware.eu/cleen2017

IEEE HPSR 2017 — IEEE Int'l. Conference on High Performance Switching and Routing, 27–30 June
Campinas, Brazil
http://hpsr2017.ieee-hpsr.org/

JULY

IEEE ISCC 2017 — IEEE Symposium on Computers and Communications, 3–6 July
Heraklion, Greece
http://www.ics.forth.gr/iscc2017/index.html

IEEE NETSOFT 2017 — IEEE Conference on Network Softwarization, 3–7 July
Bologna, Italy
http://sites.ieee.org/netsoft/

ICUFN 2017 — Int'l. Conference on Ubiquitous and Future Networks, 4–7 July
Milan, Italy
http://icufn.org/

IEEE ICME 2017 — IEEE Int'l. Conference on Multimedia and Expo, 10–14 July
Hong Kong, China
http://www.icme2017.org/

SPLITECH 2017 — Int'l. Multidisciplinary Conference on Computer and Energy Science, 12–14 July
Split, Croatia
http://splitech2017.fesb.unist.hr/

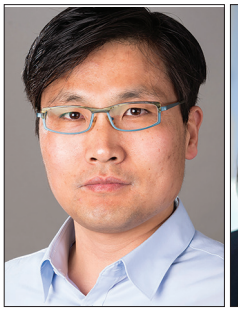
CITS 2017 — Int'l. Conference on Computer, Information and Telecommunication Systems, 21–23 July
Dalian, China
http://atc.udg.edu/CITS2017/

ICCCN 2017 — Int'l. Conference on Computer Communication and Networks, 31 July–3 Aug.
Vancouver, Canada
http://icccn.org/icccn17/

–Communications Society portfolio events appear in bold colored print.
–Communications Society technically co-sponsored conferences appear in black italic print.
–Individuals with information about upcoming conferences, Calls for Papers, meeting announcements, and meeting reports should send this information to: IEEE Communications Society, 3 Park Avenue, 17th Floor, New York, NY 10016; e-mail: p.oneill@comsoc.org; fax: + (212) 705-8996. Items submitted for publication will be included on a space-available basis.

FOG COMPUTING AND NETWORKING: PART 1


Mung Chiang


Sangtae Ha


Chih-Lin I


Fulvio Risso


Tao Zhang

Fog rises as cloud descends to be closer to the end users. Building on the foundation of past work in related areas and driven by emerging new applications and capabilities, fog computing and networking is now presenting unique opportunities to university researchers and the industry.

This Feature Topic in *IEEE Communications Magazine* consists of overview articles that span much of this growing terrain of fog. Due to a much higher volume of submissions than expected, we could only accept a small portion of the submitted manuscripts even after expanding the Feature Topic into two parts, Part 1 in April and Part 2 in August. An exciting new area often faces questions about its scope. In a separate short article immediately following this editorial, the guest editors together provide a tutorial in the form of a Q&A. Here in the rest of this editorial, we highlight the articles appearing int Part 1 of this Feature Topic.

"Optimizations and Economics of Crowdsourced Mobile Streaming" identifies the increasing demand for mobile video streaming. It is timely and interesting to read. The article proposes to use edge resources in a cooperative manner, which is to be enabled by fog computing. It opens with a descriptive section listing four types of cooperative video streaming models that pool various network resources effectively in different application scenarios. They are mobile peer-to-peer (MP2P), device-to-device (D2D), (3) bandwidth aggregation (BA), and crowdsourced mobile streaming (CMS). The authors then focus on the CMS model and introduce the corresponding optimization methods for efficient resource allocation, as well as economic incentives. Finally, the current challenges and the areas of interest in cooperative video streaming models are summarised.

Another article, "Fog-Based Transcoding for Crowdsourced Live Streaming" (CLS) looks at a very similar if not the same application in fog networks. The approach of transcoding is not a novel method, but the analysis that leads to the selection of viewers for video transcoding is. One of the noteworthy contributions of this article is that the authors propose a novel framework of a CLS system and provide experimental results to uphold their arguments. Specifically, the analysis of Twitch TV viewers' behavior and the implementation of the presented concept in a PlanetLab environment are interesting.

Fog networks face redundancy. They can leverage redundancy for robustness, and they must manage redundancy to strike trade-offs. In the article "Coding for Distributed Fog Computing," the authors provide a unifying framework for managing redundancy by optimizing coding. They illustrate the framework with two important special cases: minimum bandwidth coding and minimum latency coding. At the heart of the design choice is the trade-off between computation latency and communication load as modulated through coding.

As a continuum from cloud to things, fog physically and functionally bridges the capabilities offered in cloud and those on the edge of networks. Smartphones offer a natural point for such a bridge. In the article "RAINA: Reliability and Adaptability in Android for Fog Computing," the authors present such an architecture and zoom in on a particularly important attribute: the predictability of smartphones' service in bridging cloud and edge. The article overviews the challenges and proposes strategies for this key aspect of a smartphone-oriented fog architecture.

Cloud-RAN has been discussed as a radio access network (RAN) technology for over a decade now. Fog-RAN is now rising as an alternative for decomposing the functionalities of 5G cellular networks along the edge and into the devices. In the article "5G Radio Access Network Design with Fog Paradigm: Confluence of Communications and Computing," the authors discuss some of the key design issues, such as traffic forwarding, content caching, interworking, and security in Fog-RAN. Special attention is paid to the promise of communication and computation coming together.

Mobile edge computing (MEC) extends the cloud computing to the edge of the RAN. The article "Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges" shows the authors' vision on how context-aware collaboration among MEC servers and end-user devices can help achieve low-latency, high-bandwidth, and agile mobile services for 5G. In particular, they present three representative use cases, ranging from computation

orchestration, to collaborative video caching and processing, to interference cancellation.

In the August issue, Part 2 of this Feature Topic will appear along with an editorial introducing those papers.

BIOGRAPHIES

MUNG CHIANG (chiangm@princeton.edu) is an Arthur LeGrand Doty Professor of Electrical Engineering at Princeton University. He serves as inaugural Chairman of the Princeton Entrepreneurship Council and Director of the Keller Center for Innovation in Engineering Education. The recipient of a Waterman Award, an IEEE Tomiyasu Award, and a Guggenheim Fellowship, he works in areas such as NUM, SDP, and fog. He created the Princeton Edge Lab and co-founded the OpenFog Consortium. His MOOC reached 250,000 people, and his textbook received an ASEE Terman Award.

SANGTAE HA is an assistant professor in computer science at the University of Colorado Boulder. He received his Ph.D. in computer science from North Carolina State University. He is a co-founder and founding CTO/VP Engineering of DataMi, a mobile network startup. His research focuses on building and deploying practical network systems. He received the INFORMS ISS Design Science Award in 2014, and serves as an Associate Editor for the *IEEE Internet of Things Journal*.

CHIH-LIN I received her Ph.D.E.E. from Stanford University. She is CMCC Chief Scientist of Wireless Technologies, launched 5G R&D in 2011, and leads the C-RAN, Green, and Soft initiatives. She was on the IEEE ComSoc Board, GreenTouch EB, IEEE M&C Board Chair, and WCNC SC Founding Chair. She is on the IEEE 5G Initiative SC and Publication WG Chair, ComSoc SPC and SDB, ETSI/NFV NOC, WWRF SB, and Singapore NRF SAB. She received the *IEEE Transactions on Communications* Best Paper Award and the ComSoc Industrial Innovation Award.

FULVIO RISSO (Italy, 1971) received his B.Sc. and Ph.D. degrees from Politecnico di Torino, Italy, in 1995 and 2000, respectively. Since 2000, he has been with the Politecnico di Torino, where he is currently an associate professor and in charge of the Network and Multimedia Lab. His main areas of research interest are high-speed and flexible in-network processing, SDN, and NFV. He is an author of 90+ papers and is very active in open-source software, starting with WinPcap in 1999.

TAO ZHANG [F] joined Cisco in 2012 as the chief scientist of its Smart Connected Vehicles business. He is a cofounder and Board Director of the Open Fog Consortium, and the CIO and a member of the Board of Governors of IEEE Communications Society. He has been directing R&D for over 25 years, holds 50+ U.S. patents, and has co-authored two books, *Vehicle Safety Communications: Protocols, Security, and Privacy* (2012) and *IP-Based Next Generation Wireless Networks* (2004).

CLARIFYING FOG COMPUTING AND NETWORKING: 10 QUESTIONS AND ANSWERS

BY MUNG CHIANG, SANGTAE HA, CHIH-LIN I, FULVIO RISSO, AND TAO ZHANG

1. WHAT IS FOG COMPUTING AND HOW IS IT DIFFERENT FROM EDGE COMPUTING?

Fog computing is an end-to-end horizontal architecture that distributes computing, storage, control, and networking functions closer to users along the cloud-to-thing continuum.

The word "edge" may carry different meanings. A common usage of the term refers to the edge network as opposed to the core network, with equipment such as edge routers, base stations, and home gateways. In that sense, there are several differences between fog and edge.

First, fog is inclusive of cloud, core, metro, edge, clients, and things. The fog architecture will further enable pooling, orchestrating, managing, and securing the resources and functions distributed in the cloud, anywhere along the cloud-to-thing continuum, and on the things to support end-to-end services and applications. Second, fog seeks to realize a seamless continuum of computing services from the cloud to the things rather than treating the network edges as isolated computing platforms. Third, fog envisions a horizontal platform that will support the common fog computing functions for multiple industries and application domains, including but not limited to traditional telco services. Fourth, a dominant part of edge is mobile edge, whereas the fog computing architecture will be flexible enough to work over wireline as well as wireless networks.

2. IS FOG JUST A SMALLER CLOUD?

First, the size of the fog is flexible — it can range from a single small fog node to large fog systems comparable to existing clouds, depending on the application needs.

While fog will bring many cloud-like services closer to end users and can have smaller footprints than the cloud, it has a different vision from that of smaller or mini-clouds. Mini-clouds tend to be designed as isolated computing platforms. Fog envisions a seamlessly integrated cloud-fog-thing architecture to enable computing anywhere along the cloud-to-things continuum. Fog-to-cloud and fog-to-fog interactions will therefore be a focus of an end-to-end fog computing architecture to distribute computing functions, and then manage, pool, orchestrate, and secure the distributed resources and functions. Fog may form a hierarchical architecture between the cloud and the things, with fog nodes at different architectural levels collaborating with each other to support end-to-end applications.

Fog also concerns the control of cyber-physical systems and D2D communication, in addition to computation and storage in clouds, big or small.

3. IS FOG EQUIVALENT TO IOT?

Fog is an architecture. The Internet of Things (IoT) often refers to a set of services and applications.

An architecture decides the allocation of functionalities. It formulates and answers questions such as "who does what, and at what timescale and location?" An architecture supports many applications, some in existence today and others more futuristic.

For example, TCP/IP represents an Internet architecture. It includes several key principles, such as addressing, and allocation of functionalities, such as congestion-independent, hop-by-hop routing, and congestion-dependent, end-to-end session control. Applications that leverage TCP/IP have come from a wide and increasing range: from the web to emails and from P2P to video streaming.

The relationship between fog and IoT is similar to that between the Internet architecture and the web applications. Fog also supports other areas of applications, such as those in fifth generation (5G) cellular or embedded artificial intelligence.

4. IS FOG FOR COMPUTATION, OR COMMUNICATION, OR CONTROL?

Fog is an umbrella term that includes an architecture for computation, an architecture for communication, an architecture for storage, and an architecture for control (both control of the network itself and networked control in cyber-physical systems).

For example, fog computing explores new ways to decompose a computational task so as to match an underlying computation substrate that is heterogeneous (in hardware and software capabilities), volatile (in availability, mobility, and security), and constrained (by bandwidth or battery). Fog communication explores how devices may talk to each other despite intermittent global connectivity. Fog control explores how clients might crowd-sense network conditions and self-configure, and how to leverage small and almost deterministic latency to enable feedback control loops.

5. WHAT ARE THE UNIQUE ADVANTAGES OFFERED BY FOG?

Unique advantages that are potentially offered by fog can be summarized with an acronym: "SCALE." These advantages in turn enable new services and business models, and may help broaden revenues, reduce cost, or accelerate product rollouts.

Security: While fog faces unique security challenges, it also offers certain advantages. In particular, by reducing the distance that information needs to traverse, there is less chance of eavesdropping. By leveraging proximity-based authentication challenges, identity verification can be strengthened.

Cognition: Awareness of client-centric objectives. A fog architecture, aware of customer requirements, can best determine where to carry out the computing, storage, and control functions along the cloud-to-thing continuum. Fog applications, being close to the end users, can be built to be better aware of and closely reflect customer requirements.

Agility: Rapid innovation and affordable scaling. It is usually much faster and cheaper to experiment with client and edge devices rather than waiting for vendors of large network and cloud boxes to initiate or adopt an innovation. Fog will make it easier to create an open marketplace for individuals and small teams to use open application programming interfaces (APIs), open software development kits (SDKs), and the proliferation of mobile devices to innovate, develop, deploy, and operate new services.

Latency: Real-time processing and cyber-physical system control. Fog enables data analytics at the network edge and can support time-sensitive control functions for local cyber-physical systems. This is essential for not only commercial applications but also for the Tactile Internet vision to enable embedded AI applications with millisecond reaction times.

Efficiency: Pooling resources along the cloud-to-thing continuum. Fog can distribute computing, storage, and control functions anywhere between the cloud and the endpoint to take full advantage of the resources available along this continuum. It can also allow applications to leverage otherwise idle computing, storage, and networking resources abundantly available on network edge and end-user devices such as tablets, laptops, smart home appliances, connected vehicles and trains, and network edge routers. Fog's closer proximity to the endpoints will enable it to be more closely integrated with end-user systems to enhance overall system efficiency and per-

formance. This is especially important for performance-critical cyber-physical systems.

6. Is Fog Good or Bad for Security and Privacy?

Fog systems and applications will often be distributed and operated remotely. Some fog systems can also be resource-constrained. Compared to centralized clouds, such distributed, remote, and resource-constrained fog systems pose additional security challenges often encountered in distributed systems. On the other hand, fog can bring more processing resources closer to the endpoints to help better protect the vast population of diverse endpoints that often do not have sufficient resources to adequately protect themselves. In other words, fog systems can provide a wide range of local security services to make the IoT as a whole more secure. For example, fog systems can perform local security monitoring, local threat detection, and local threat protection functions on behalf of the endpoints. Fog nodes can also serve as proxies of the endpoints to help manage and update the security credentials and software on the endpoints, eliminating the often impractical needs for all the endpoints to directly communicate with the remote cloud for such functions.

7. Will the Need for Fog Diminish as Network Capacity and Delay Improve over Time?

While it is true that a primary benefit of fog computing is its ability to reduce latency and delay, the drivers for fog go far beyond pure latency issues to include a variety of operational, regulatory, business, and reliability issues.

For example, instead of the traditional way of adding new applications by adding dedicated new local servers and networking gear, fog can provide a common end-to-end platform for all services provided to each customer. This can provide a unified platform to support life cycle management, networking, and security for all applications, which will reduce system complexity and costs and also allow applications from different providers to better interact with each other rather than stay siloed on their dedicated hardware and software platforms. Fog can enable critical services to be operated autonomously or managed from the cloud, the perimeter, or a variety of points in the network. Fog is equally advantageous for areas where network connectivity can be unreliable due to weather or other conditions. It can also significantly reduce network bandwidth loads through its proximity to where the data is generated. With fog, local operational and business policies can be applied to enable more efficient local data processing and analytics on premises.

As another example in cellular networks, cloud RAN, with centralized or distributed network architecture, has the advantage of being physically close to the end users and the capability of utilizing the network resources at the edge. Consequently, the cloud radio access network (C-RAN) will be an integral part of the solution to meet stringent network delay requirements that the traditional RAN network may fail to meet. Consequently, the Third Generation Partnership Project (3GPP) is now discussing a RAN architecture that contains both the central units and the distributed units. Extending prior notions in C-RAN, fog network is unique in the sense that the end user computing and storage resource is considered as an integral part of the whole network, by forming ad hoc subnets among end nodes. Careful exploitation of such features will bring unique values for fog networks.

8. What New Technologies and Standards, If Any, Do We Need to Develop for Fog?

Fog systems will need to interact with each other, with the clouds, and with a diverse range of user end devices. Therefore, the success and wide adoption of fog computing will rely on standards. While fog computing can benefit from many existing standards, new standards may also be required, for example, in the following areas:

Building unified fog-cloud platforms: Interfaces and protocols for the fog and the cloud to interact with each other to enable unified cloud-fog service platform and applications, move computing functions and applications between the cloud and the fog, pool resources distributed in the cloud and the fog, and manage the life cycle of the fog systems and applications.

Support distributed and hierarchical fog systems over possibly heterogeneous, volatile, and constrained physical resources: interfaces and protocols for different hierarchical levels in a fog system to interact with each other, and for different fog systems at the same hierarchical level to collaborate with each other to serve as each other's backup.

Access to fog services: A fog system, bringing resources closer to end users, can enable a wide range of new fog-based services. Standards will be required for users and their devices to interact with the fog system to discover, request, and receive fog services. So will automatic and lightweight bidding mechanisms for access to fog resources and services to reinforce the economic sustainability of the fog computing model, and enabling economic transactions.

Data management: Local processing and management of data is one of the important drivers for fog computing. Data, however, comes from an increasingly wide range of sources. Data management also imposes widely diverse requirements from industry to industry. New standards may be required to manage the diverse data, such as storing, accessing, and securing the data distributed in the fog and cloud.

Security and privacy: A distributed and remotely operated fog system can pose new security challenges not present in centralized systems. Addressing these new challenges may require new standards. For example, fog computing will need to run a diverse set of local hardware platforms. Therefore, new interfaces may be required for fog software to interact with the various hardware platforms, which may be provided by different vendors, to ensure a trusted computing environment. New interfaces and protocols may be required for automatic detection of security compromises in a distributed and remote fog system, and also for remote and automatic responses to security compromises.

Furthermore, although standards may exist for some fog computing needs, additional requirements in fog computing environments (e.g., low-latency, large number of resource-constrained devices) may necessitate new standards that are more suitable for fog computing environments.

9. What New Research Challenges Do We Have to Address to Enable Fog?

Research challenges in fog span a wide range: from computation decomposition over heterogeneous and constrained nodes to cloud-fog interface definition, from state consistency in dispersive computing to elastic storage over volatile substrate, from pricing for economic incentives to scalable security measures. Fundamental to these topics is the intrinsic trade-off between "local" and "global" and between "brick" and "click" as we slide between cloud and things in deciding where to allocate a function and how to glue them back together.

For example, fog computing enables a complex service to possibly be delivered through a set of elementary software elements that operate on heterogeneous nodes, such as end user terminals and local servers, but also network elements and data centers. The problem of the orchestration of

the above complex services is definitely an important challenge, complicated by the highly dynamic environment, the many fog-enabled applications installed on end user devices and things, and the necessity to support different administrative domains, to adapt the service to the extreme heterogeneity of the infrastructure, and to adapt the service (and the orchestration algorithms) to the external environment. For instance, fog applications cannot always count on the availability of powerful computing devices that can execute complex orchestration algorithms; in critical conditions (e.g., broken infrastructure as in the case of an earthquake), the fog infrastructure has to be able to orchestrate services even in the presence of limited computing capabilities, with an intrinsic degree of resiliency. Along this line, another important challenge is the capability to create self-adapting applications, which are able to automatically adapt their behavior based on the surrounding environment, for example, in case some required services (e.g., high-capacity storage or a high-precision sensor) cannot be reached, while still being able to deliver the service the user is expecting, albeit with some degradation.

10. What Commercial Opportunities Will Fog Bring?

Fog computing will bring many new commercial opportunities and will disrupt the existing industry landscapes and business models, disrupting the balance of power along the industry food chain.

For example, networking functions (e.g., routing and switching), application servers, and storage functions are already converging into integrated "fog nodes": edge devices that integrate edge router and local application server and storage functions are already commercially available. The emerging fog systems will empower the cloud to do what it cannot effectively do today by, for example, acting as proxies to connect and then provide cloud services to the many devices that cannot be practically connected to the cloud directly. A growing range of innovative fog-based services, including fog systems and services as a service, will emerge. The cloud and the fog will converge into unified end-to-end platforms and provide integrated services and applications, creating opportunities for fundamental disruptions to the existing cloud computing business models. Players of all sizes will be able to deploy fog systems and operate fog services. And the list goes on.

Optimizations and Economics of Crowdsourced Mobile Streaming

Ming Tang, Lin Gao, Haitian Pang, Jianwei Huang, and Lifeng Sun

ABSTRACT

Mobile video traffic accounts for more than half of the global mobile data traffic nowadays, and the ratio is expected to further increase in the near future. However, providing high quality of experience for video streaming in mobile networks is challenging due to the heterogeneous and varying wireless channel conditions. To meet the increasing demand of high-quality mobile video streaming services, researchers have proposed several cooperative video streaming models that enable mobile users to download video contents cooperatively. The key idea is to pool network edge resources so as to either alleviate the load on the video servers and the cellular network, or alleviate the impact of channel variations and improve resource utilization. In this article, we review four types of cooperative video streaming models that pool various network resources effectively in different application scenarios. Then we focus on the crowdsourced mobile streaming model, which aims to pool users' download capacities in order to alleviate the impact of channel variations and achieve efficient utilization of network resources. We introduce the corresponding optimization issue of efficient resource allocation and the economic issue of user cooperation. We also outline future challenges and open issues in cooperative video streaming models.

INTRODUCTION

With the development of mobile networks and mobile devices, users now are capable of enjoying video streaming services over mobile networks. Cisco reported, in February 2016, that the mobile video traffic already accounted for 55 percent of total mobile traffic, and it is expected to grow at an annual rate of 62 percent in the next few years [1]. The heavy video traffic challenges mobile network infrastructure and video servers. Compared to users in wired networks, mobile users experience heterogeneous and time-varying channel conditions and network resources, in the sense that different users will have different achievable data rates depending on, for example, their network operators and locations. The heterogeneity and variation induce challenges for providing high-quality, stable, and smooth mobile video streaming experiences to mobile users.

To address the challenges, adaptive bit rate (ABR) streaming [2] has been proposed and widely used for wireless video streaming. Recent standards and commercial instances of ABR include MPEG Dynamic Adaptive Streaming over HTTP, Apple HTTP Live Streaming, and Microsoft Smooth Streaming. In ABR, a video source is partitioned into multiple small video pieces, called segments, and each segment is encoded at multiple bit rates. Mobile users can choose the bit rate of each segment and hence can dynamically adapt their videos to their heterogeneous and time-varying network conditions.

Through a proper bit rate adaptation method (i.e., how a user should choose the bit rate of each segment), ABR enables video streaming services to better adapt and utilize a user's wireless resources. The user's quality of experience (QoE), however, is still restricted by the video server bandwidth and the user's own channel conditions. The adoption of cloud computing may alleviate the load on the video server through offloading upload bandwidth to the cloud. However, cloud-based video streaming may further add additional delay in the system, and does not help resolve the limitation due to the user's own channel conditions. On the other hand, through edge resource pooling in the framework of fog computing, one can effectively integrate the communication resources of multiple edge devices and exploit the diversity of users' channel conditions. Such edge resource pooling can reduce video server load (through letting edge devices serve video users with their available contents) as well as increase network reliability, flexibility, and efficiency, which is especially useful for mobile networks with heterogeneous and varying channels.

Inspired by these ideas, researchers have proposed several models for *cooperative video streaming:*

- *Mobile peer-to-peer (MP2P) model* [3], where video users partially fulfill the role of servers by forwarding their downloaded videos to other video users through the Internet
- *Device-to-device (D2D) model* [4], where video users exchange their downloaded video segments to nearby video users through short-distance D2D wireless links
- *Bandwidth aggregation (BA) model* [5], where a video user and his/her nearby idle users pool their network resources for this single video user's streaming need

Mobile video traffic accounts for more than half of the global mobile data traffic nowadays, and the ratio is expected to further increase in the near future. However, providing high quality of experience for video streaming in mobile networks is challenging due to the heterogeneous and varying wireless channel conditions.

Ming Tang and Jianwei Huang (co-corresponding author)} are with The Chinese University of Hong Kong; Lin Gao is with Harbin Institute of Technology; Haitian Pang and Lifeng Sun (co-corresponding author) are with Tsinghua University.

Digital Object Identifier: 10.1109/MCOM.2017.1600830

The MP2P model is the peer-to-peer (P2P) model applied in mobile networks. In MP2P, mobile users with some downloaded segments can forward those segments to other users in need, to partially fulfill the role of a video server.

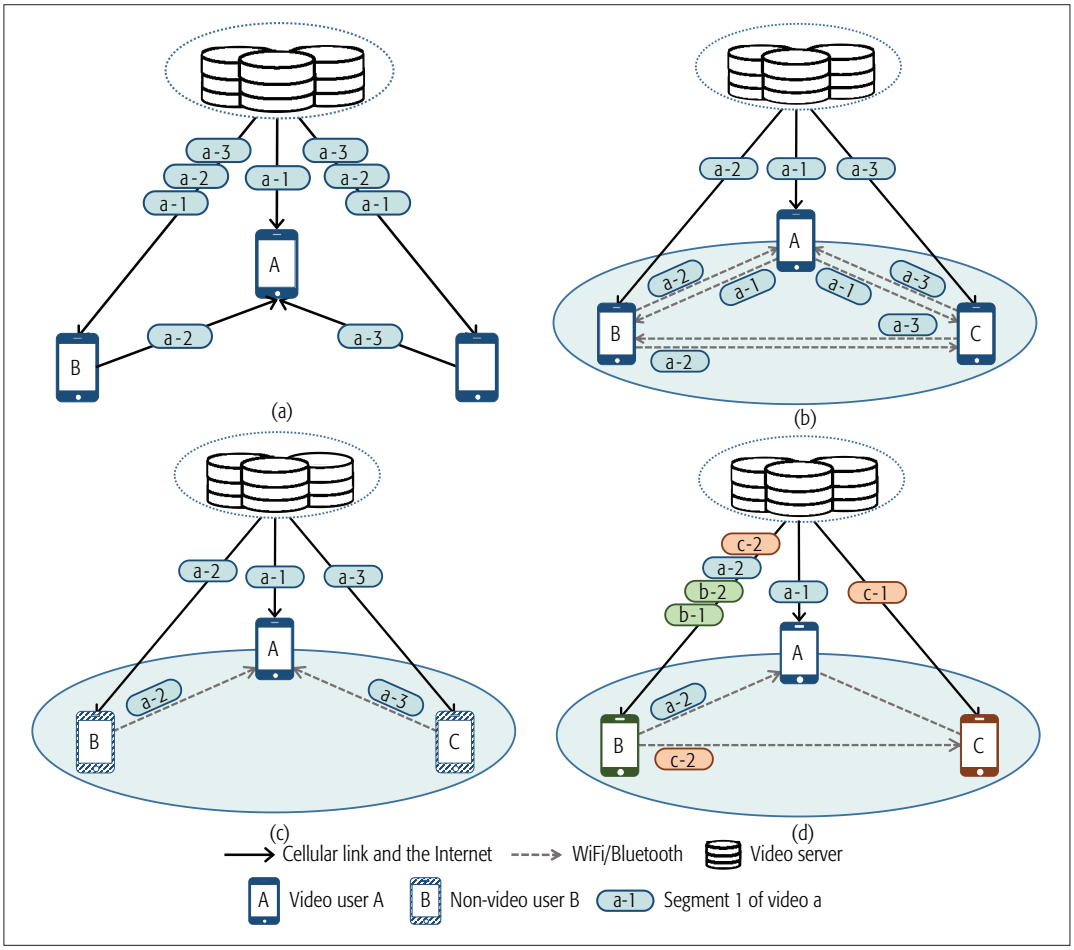


Figure 1. Cooperative video streaming: a) MP2P model; b) D2D model; c) BA model; d) CMS model.

- *Crowdsourced mobile streaming (CMS) model* [6–8], where nearby video users that watch different videos pool their network resources to download the videos

In these models, the MP2P model pools peers' uplink network capacities to the Internet to alleviate server load; the D2D model pools downloaded segments to alleviate server and cellular network load; and the and BA and CMS models pool downloaded network resources to increase resource efficiency and alleviate the impact of channel variations.

In order to effectively support these cooperative models, research needs to address three key issues:

- *Technical issues:* How are real-world cooperative video streaming systems constructed, including designing the cooperative structures and managing the channel interferences due to cooperations?
- *Optimization issues:* How is the video streaming operation scheduled among multiple users, including bit rate adaptation and resource allocation?
- *Economic issues:* How are incentive mechanisms designed to motivate users to share resources cooperatively?

In this article, we provide a comprehensive understanding of the cooperative video streaming model so as to illuminate the key ideas, challenges, and possible solutions for the edge resource pooling approach in fog computing. We first provide an overview of several cooperative video streaming models and a further introduction on the key issues. We then focus on the CMS model, as it considers the most complicated scenario in which different video users watch different videos. We discuss both offline and online scheduling algorithms for achieving efficient or nearly efficient resource allocation, and describe a truthful incentive mechanism that effectively motivates user cooperation. Finally, we outline some future challenges and open issues for cooperative video streaming.

Cooperative Video Streaming

Overview of Cooperative Video Streaming

Cooperative video streaming enables mobile users to cooperate and share their wireless links or downloaded resources in order to enhance the video streaming experiences of some or all of the users. In this section, we introduce four types of cooperative streaming models and compare their key features.

The MP2P model is the P2P model applied in mobile networks. In MP2P, mobile users with some downloaded segments can forward those segments to other users in need to partially fulfill the role of a video server. Figure 1a shows an example of the MP2P model: user A downloads segment 1 of *video a* directly from the server, and downloads segments 2 and 3 of *video a* from users B and C, respectively. In the D2D model, mobile users share their downloaded segments with nearby mobile users through D2D wireless links, such as WiFi or Bluetooth. As shown in Fig. 1b, the three users download segments from the

Models		MP2P model [3]	D2D model [4]	BA model [5]	CMS model [6–8]
Pooled resources		Upload capacity	Downloaded segments	Download capacity	Download capacity
Key objective		Alleviate server congestion	Alleviate server and cellular load	Increase resource efficiency	Alleviate channel variations; increase resource efficiency
Scenario	Interaction	Internet	Local	Local	Local
	Video session	Multiple	Multiple	One	Multiple
	Video number	One	One	One	Multiple

Table 1. Model comparisons.

severs and share them through D2D links. In the BA model, a video user and his/her nearby idle users pool their network resources for the single video user's streaming. For example, in Fig. 1c, user A is the only user who watches a video. Users B and C download segments 2 and 3 for user A, respectively, and forward them to him/her through D2D links. In the CMS model, mobile video users pool their network resources to satisfy all the users' different video streaming needs. Such a model takes care of the load balancing issue among mobile users naturally, as it intends to properly allocate the network resources among mobile users to maximize social welfare. Figure 1d shows an example where three users watch different videos. User B has a better downlink channel, so he/she not only downloads two segments of *video b* for him/herself, but also downloads one segment of *video a* and one segment of *video c* to satisfy the needs of users A and C, respectively.

Table 1 provides further comparisons among these models in other dimensions: interaction, whether the cooperation happens among remote users via the Internet or local users; video session, how many users watch videos; and video number, how many videos the users watch (if more than one, different users may watch different videos).

Key Issues

Here we discuss three key issues of cooperative video streaming in a bit more detail.

The first type are *technical issues* in constructing these cooperative video streaming models. First, how do we enable cooperative structures? For the MP2P model, although wired P2P structure can be implemented in MP2P, MP2P has to address new challenges: varying wireless channel (mostly caused by device mobility) and limited device storage capacity. The varying channels make it hard for uploading users to maintain stable upload speeds, and the limited device storage capacity makes it unreasonable to store a large number of segments in mobile devices. For the other three models, the key challenges include how to discover and establish D2D connections with limited device energy capacity, and how to enable simultaneous data transmission and reception through multiple interfaces (e.g., downlink cellular interface and D2D interface). There have been several recent efforts in designing cooperative structures to overcome these challenges, such as [5] for the BA model. Second, how do we manage the interference in the cooperative frameworks, for example, the interference between cellular and D2D links as well as the interference among D2D links themselves? This issue is most relevant in the context of D2D communications; so many existing proposals (e.g., [9]) can be implemented. For example, spectrum splitting, which separates the spectrum usage for cellular and D2D links, can be used for handling the interference between cellular and D2D links, and power control and radio resource allocation, which optimize the power and spectrum allocation, respectively, among multiple D2D pairs, can be used for handling the interference among D2D links.

The second type are *optimization issues* on bit rate adaptation — how video users select the bit rates of their segments to enhance their video quality of experience — and resource allocation — how the network resources should be allocated to achieve certain objectives, such as social welfare maximization. The common key challenge for addressing both issues is the asynchronous downloading operation among users. Cao *et al.* [4] studied a D2D group formation problem for scheduling video segments among users in the D2D model. Lin *et al.* [6] attempted to overcome this issue by studying a virtual synchronous downloading operation for understanding the bit rate adaptation and the resource allocation in the asynchronous CMS model. However, completely and effectively addressing the asynchronous operation is still an open issue.

The third type are *economic issues* on incentive mechanism design. Sharing resources is always costly, especially for mobile users who have limited communication, storage, and battery resources. Hence, we need an effective incentive mechanism to motivate users to share and cooperate. A key challenge for incentive mechanism design is private user information. The issue is particularly complicated in cooperative video streaming because the video segments are encoded at multiple bit rates, and users can have different private valuations for the segment encoded at each bit rate. Kang *et al.* [3] proposed a credit-based incentive mechanism for MP2P streaming, with the goal of jointly maximizing the revenue of the helper and the utility of the help receiver. The authors in [3] focused on the uploading and downloading bandwidth allocation, while ignoring the situation of multi-bit-rate encoded streaming. Ming *et al.* [7] proposed a truthful auction mechanism in the CMS model, with the goal of maximizing social welfare through proper resource allocation and bit rate adaptation, simultaneously providing sufficient motivations for the helper.

Crowdsourced Mobile Streaming

Among the four cooperative models, the CMS model focuses on the most general (and arguably most commonly encountered) scenario in which

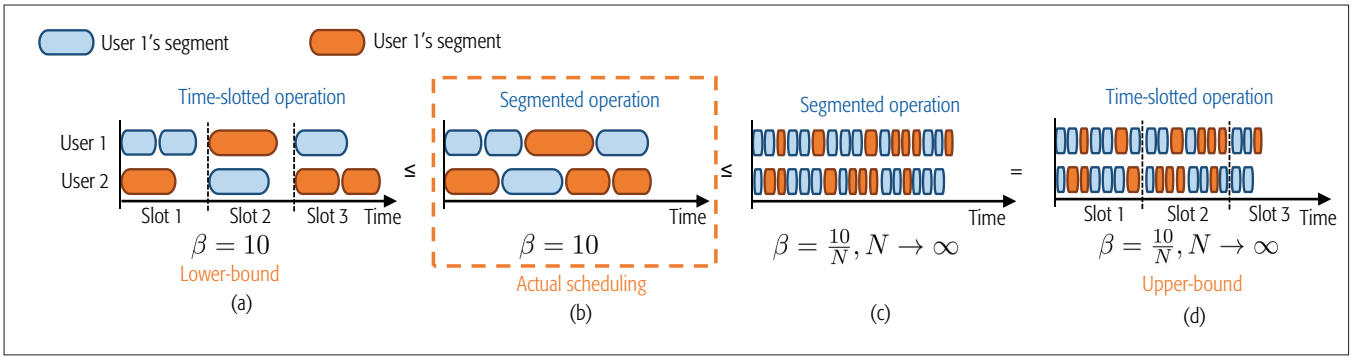


Figure 2. Upper-bound and lower-bound of the maximum social welfare of the segmented operation.

different users watch different videos. In this model, mobile users not only download videos asynchronously, as in the BA and D2D models, but also request and watch videos asynchronously from different video servers. Hence, it is challenging to properly schedule the asynchronous cooperation and satisfy all the users' heterogeneous requirements. It can also be difficult to accurately evaluate users' contributions for helping other users with different videos and bit rate requirements. In this section, we concentrate on the CMS model, discussing its optimization and economic issues.

In the CMS model, mobile users pool their resources for more effective video streaming. Through effective allocation of the network resources of all the users, the crowdsourced framework can reduce the impact of cellular link variations at individual levels and exploit the positive network effects (i.e., users' heterogeneous cellular links and video requirements). Social welfare in the CMS model is the total welfare achieved by all users, and is defined as the difference between the users' QoE and the users' cost. The QoE depends on video qualities, rebufferings, and quality degradations, while the cost depends on energy consumption and cellular data payment [10].

Next we introduce solutions to optimization issues and economic issues in the CMS model. For the optimization issues, we aim to properly schedule the cooperation and bit rate adaptation to maximize social welfare in an offline scheduling operation benchmark and online scheduling operation. For each video user, we need to decide when and for whom he/she is going to download segments at what bit rates. For the economic issues, we aim to design an incentive mechanism that can motivate a user to truthfully reveal the user's private information and achieve social welfare maximization. For each video user, we need to decide how much he/she should be paid when he/she downloads video segments for different users at different bit rates. Note that although both issues involve the social welfare maximization problem, to resolve the economic issue we need to handle the private user information, such as buffer size and bit rate preference, while the optimization issue focuses on social welfare maximization under the assumption of complete (or public) user information. In distributed scheduling, as in the online scheduling optimization section, the public user information can be obtained through information exchange among nearby users.

Offline Scheduling Optimization

Offline scheduling optimization, as a benchmark, aims at maximizing social welfare assuming complete (or public) network information and user information. Here, the network information refers to all the users' historical and future cellular link capacities, and such information is known to all users publicly. In the offline case, we discuss the theoretical *social welfare performance bound* of the proposed crowdsourced system, which serves as a benchmark for the online scheduling solutions.

Directly solving the social welfare maximization problem even in the offline case is challenging. First, the video downloading of each user involves *segmented operation* in ABR streaming. Specifically, as illustrated in Fig. 2b, two users download new segments in an asynchronous manner, where the downloading time of a segment depends on the amount of data and the channel condition of the download. Second, social welfare optimization is a mixed-integer program, containing both discrete variables (e.g., bit rate and user set) and continuous variables (e.g., downloading start time), and has integral operations (resulting from calculating the download volumes of the varying cellular channels). These features make solving the problem challenging.

To understand the offline scheduling, in [6] we proposed a virtual *time-slotted operation*, under which we can characterize the upper bound and lower bound of the maximum social welfare of the original asynchronous operation. In the virtual time-slotted operation, as in Fig. 2a, two users schedule download segments slot by slot in a partially synchronized fashion. With this virtual operation, we can focus on the segment downloading of each user in each time slot: how many segments to download, for which users, and at what bit rates. The optimization problem in the time-slotted operation can be formulated as a linear integer programming problem, and can be solved by many classic methods.

We can show that the performance of the segmented operation is upper- and lower-bounded by the time-slotted operation with proper system parameter choices. Let β be the segment length in terms of playback time. The *lower bound* of the maximum social welfare of the segmented operation (Fig. 2b) is the time-slotted system with the same video segment length (Fig. 2a), because the downloading operation under the time-slotted operation is feasible under the segmented operation, but not vice versa. The *upper bound* of the maximum social welfare of the segmented operation (Fig. 2b) is the time-slotted system with the segment length approaching zero (i.e., Fig. 2d),

through dividing an integral value N that approaches infinity. Specifically, considering the segmented operation, the social welfare will be non-decreasing when the segment length β decreases (comparing Figs. 2b and 2c), because the downloading operation under the larger segment length is still feasible when the segment length decreases, but not vice versa. Moreover, when segment length approaches zero, the operation under time-slotted operation can be equivalently achieved under the segmented operation (comparing Figs. 2c and 2d). Therefore, we can obtain the upper-bound and lower-bound social welfare of the asynchronous segmented operation through calculating the social welfare of the time-slotted operation with different choices of segment lengths.

Online Scheduling Optimization

In practice, however, network condition varies randomly over time, so it is difficult to obtain future and global network information, as assumed in offline scheduling. This motivates us to study the practical online scheduling problem, where the network information is incomplete (i.e., only historical and current network information is available). Note that the user information is still publicly known. The key question is *how do we schedule the segment downloading among multiple users and choose the bit rate of each segment in order to maximize the (expected) social welfare, considering the uncertain and stochastically changing future network information?*

We propose an online scheduling algorithm based on the Lyapunov optimization [11] framework: When a user is ready to download a new segment, he/she will decide on the segment receiver and the segment bit rate in order to minimize an objective function named *drift-plus-penalty*, which corresponds to all the users' buffer changes minus the social welfare times an adjustable coefficient. Intuitively, the objective is to enhance the social welfare while balancing the users' buffer sizes. Consideration of the buffers will help avoid packet drops caused by buffer overflows and video freezing due to empty buffers. Consideration of social welfare will incorporate various factors that affect users' QoE, such as bit rate satisfaction and bit rate fluctuation loss, and downloading and transmitting cost. We show that this algorithm converges to the theoretical performance bound of the offline scheduling system asymptotically, with an approximation error bound that is controllable through the adjustable coefficient mentioned above.

We test the performance of the online algorithm through numerical examples of 50 video users. In each simulation run, each user has a randomly generated cellular link capacity simulated based on real-world data traces (provided by BesTV, an over-the-top video service provider in China) that correspond to the given average link capacity. Figure 3 shows the comparisons of average social welfare among the Lyapunonv-based online algorithm, the bandwidth-based adaptation algorithm [12] (in which receiver and bit rate are chosen according to the downloader's bandwidth), and buffer-based adaptation algorithm [13] (in which receiver and bit rate are chosen according to all users' buffer sizes). We also compare the cooperation scenario (users cooperate based on the CMS model) and the noncooperation scenario (users do not cooperate). The numerical results in Fig. 3 suggest that cooperation always increases the average social welfare compared to the noncooperative case, and our proposed Lyapunov-based algorithm always has the largest average social welfare compared to the other two benchmark algorithms.

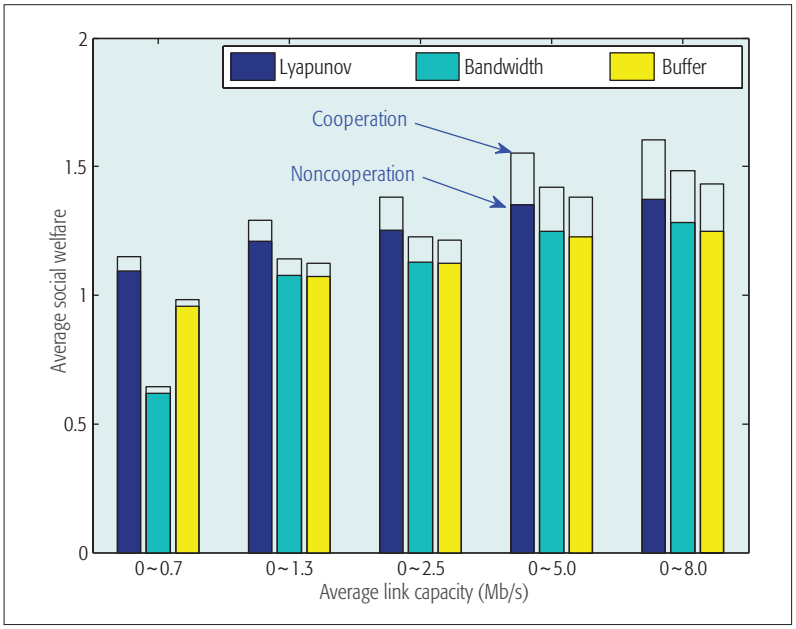


Figure 3. Social welfare comparisons among the Lyapunov-based algorithm, bandwidth-based algorithm, and buffer-based algorithm; and between cooperation and noncooperation.

Economic Incentive Mechanism

Providing help to other users can be costly, so mobile users may not be willing to participate in CMS to share their network resources. Hence, we propose an incentive mechanism to motivate user cooperation. In this section, the network information is incomplete, and user information is private. The key question is *for each segment to be downloaded, who is the segment receiver, what is the segment bit rate, and how much should the receiver compensate the downloader?*

The incentive mechanism design is challenging in the CMS model due to users' private valuations on multi-bit-rate encoded segments. Specifically, a user's preference for segment bit rate and his/her corresponding valuation is his/her private information and may vary over time. The diverse and varying private valuation induces difficulties in evaluating downloaders' contributions to cooperation and determining the proper incentive levels.

To handle the issue of private valuations, in [7] we proposed an auction-based mechanism for the CMS model: When a user is ready to download a segment, he/she will initiate an auction to determine the segment receiver, the segment bit rate, and the payment. The key here is that each bidder has to specify a multidimensional bid on the segment to be downloaded, consisting of his/her intended segment bit rate and the price he/she is willing to pay. This motivates us to consider a multidimensional auction [14].

Figure 4 illustrates an example of a second-score multidimensional auction mechanism. First, each bidder (potential receiver) submits a two-dimensional bid, consisting of the *intended*

When a user is ready to download a segment, he will initiate an auction to determine the segment receiver, the segment bitrate, and the payment. The key here is that each bidder has to specify multidimensional bid on the segment to be downloaded, consisting of his intended segment bitrate and the price he is willing to pay. This motivates us to consider a multidimensional auction.

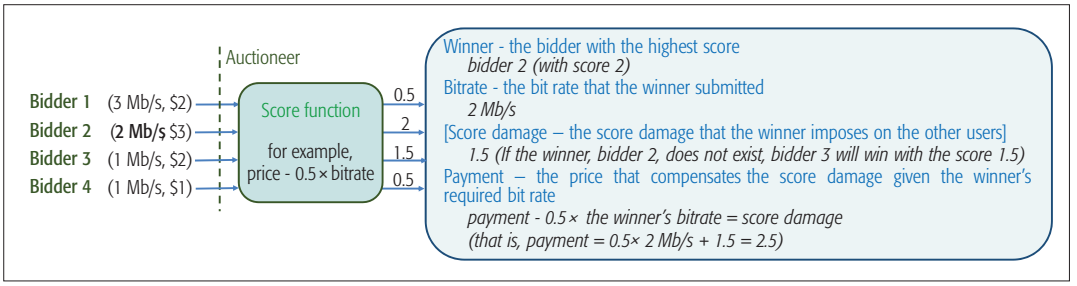


Figure 4. An example: second-score multidimensional auction.

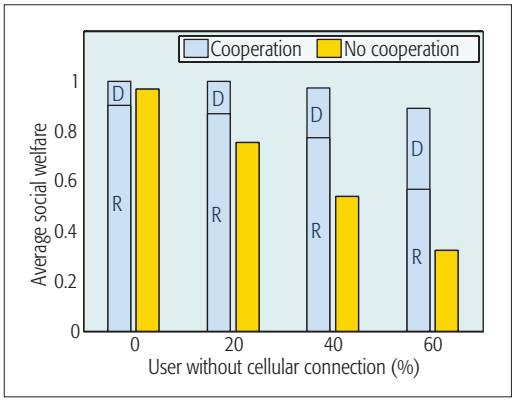


Figure 5. Social welfare comparison between cooperation and noncooperation. D: welfare obtained through downloading segments; R: welfare obtained through receiving segments.

bit rate and the *intended price* for the segment to be downloaded by the downloader. Second, the auctioneer transforms the two-dimensional bids into one-dimensional values through an additive score function (i.e., the price minus an increasing function of the bit rate). Finally, the auctioneer determines the auction results — the winner, the bit rate of the segment to be downloaded, and the payment — based on the second score rule. In such an auction, we show that the bidders always truthfully reveal their valuations through submitting their intended prices. Moreover, the score function can be flexibly chosen to achieve certain objectives, such as social welfare maximization or the auctioneer's payoff maximization. For example, when the score function is defined as the submitted price minus the cost of downloading the segment with the submitted bit rate, this auction mechanism will maximize social welfare.

Next we show the performance of the proposed auction mechanism in numerical examples with 50 users. Each user wants to watch a 50-s video, and some of the users are disconnected to the Internet. (In each simulation run, each user has a randomly generated cellular link capacity simulated based on real-world data traces provided by BesTV.) We compare the social welfare between a cooperation scenario (users cooperate based on the CMS model) and a noncooperation scenario (users do not cooperate). In the cooperation scenario, we also show users' average welfare obtained through downloading segments (denoted by *D*), that is, *as a downloader, the user's received compensation (payment) minus his/her cost of downloading*, and users' average welfare obtained through receiving segments (denoted by *R*), that is, *as a receiver, the user's utility achieved due to video consumption minus his/her payment to the downloaders*. Note that a user may act as both downloader and receiver at different time instances, so a user's welfare contains both downloading welfare and receiving welfare. The sum of all the users' downloading welfare and receiving welfare is the social welfare. Figure 5 shows that as the percentage of users without the Internet increases, the social welfare in the noncooperation scenario decreases dramatically, while the social welfare in the cooperation scenario is relatively stable, and the decrease is only 10.9 percent when the percentage of disconnected users changes from 0 to 60 percent. Moreover, as the percentage of users without the Internet increases, the welfare achieved through downloading increases due to the fact that more users require help, and the welfare achieved through receiving decreases due to the reduction of network resources and increased competition.

Moreover, to reduce the overhead due to the frequently initiated auction, we also proposed an incentive mechanism that addresses multi-object segment allocation. Due to space limits, we refer readers to [8] for details.

Demonstration System

We further constructed a demo system using Raspberry PI Model B+ (https://www.raspberrypi.org/) [8]. In the demo system, Raspberry PIs represent the mobile devices, each of which is equipped with an LTE USB modem for LTE connections and a WLAN adapter for WiFi connections. The demo system can support dynamic group joining and leaving through UDP broadcasting, so there is no need for centralized control. After forming a group, the mobile devices cooperatively download video segments via LTE, send signaling messages and forward video segments to other devices (if needed) through TCP transmissions. The cooperation is scheduled using our proposed incentive mechanism. Experiments over the demo system showed that the additional latency caused due to the auction-based resource allocation mechanism is 100 ms per auction. In practice, the length of a video segment is often 2, 5, or 10 s; hence, the implementation overhead of the auction mechanism is between 1 and 5 percent.

Future Challenges and Open Issues

In spite of recent efforts on addressing technical, optimization, and economic issues, there are still many future challenges and open issues for cooperative video streaming.

Human mobility makes user cooperation groups time-varying, hence making it harder to

schedule resources effectively. Frequent disconnections of communications among users can significantly increase the signaling overhead of the cooperation. For example, a helper may find a receiver disconnected after downloading the requested segment. Thus, it is important to design an effective and robust scheduling algorithm by taking into consideration the uncertainty introduced by user mobility.

Social relationship reflects each user's reputation and preference over cooperation. Specifically, a user with a better reputation in previous cooperation may attract more helpers, and friends in the real world tend to cooperate since they are familiar with each other. The consideration of social relationship can make the incentive mechanism more effective.

Security and privacy are always crucial in wireless networks, especially when mobile users share local network resources and individual information frequently. It is important to design proper authentication and monitoring schemes, which should be designed to support real-time video streaming services together with distributed and massive D2D connections.

Interventions of content providers and network operators: Most researchers mainly focus on the cooperation among video users, without considering the potential involvement of network operators and video content providers. In practice, network operators may be reluctant to support the crowdsourced networking scheme among users, and some network operators (e.g., AT&T in the United States) have started to charge additional fees for "tethering" among users. Considering such an intervention, Meng *et al.* [15] provide an initial study on deriving the optimal data and tethering price for the crowdsourced networking scheme. Moreover, video content providers may not be willing to support user cooperation, for example, when a user with a certain monthly subscription for an unlimited video plan downloads video for another user with a usage-based video subscription plan. Hence, to achieve the cooperation of users, we need to consider not only the incentives for users but also the incentives for the content providers and network operators.

Conclusion

Cooperative video streaming promotes mobile video streaming services by enabling mobile users to provide services (or resources) to each other or enabling mobile users to pool their network resources. In this article, we introduce four types of cooperative models and discuss key issues for model implementation. We further concentrate on the CMS model, which is applicable for a general scenario in which different users watch different videos, and introduce the optimization and economic issues and solutions in this model. We also outline some future challenges and open issues on which researchers can further work.

Acknowledgment

This work is supported by the General Research Funds (Project Numbers CUHK 14202814, 14206315, and 14219016) established under the University Grant Committee of the Hong Kong Special Administrative Region, China, and the NSFC (Grant Number 61472204 and 61521002).

References

[1] Cisco VNI, Global Mobile Data Traffic Forecast Update, 2015–2020, white paper, accessed Dec. 20, 2016.
[2] S. Akhshabi, A.C. Begen, and C. Dovrolis, "An Experimental Evaluation of Rate-Adaptation Algorithms in Adaptive Streaming over HTTP," *Proc. ACM MMSys*, 2011, pp. 157–68.
[3] X. Kang and Y. Wu, "Incentive Mechanism Design for Heterogeneous Peer-to-Peer Networks: A Stackelberg Game Approach," *IEEE Trans. Mobile Comp.*, vol. 14, no. 5, May 2015, pp. 1018–30.
[4] Y. Cao *et al.*, "SoCast: Social Ties Based Cooperative Video Multicast," *Proc. IEEE INFOCOM*, 2014, pp. 415–23.
[5] T. V. Seenivasan and M. Claypool, "CStream: Neighborhood Bandwidth Aggregation For Better Video Streaming," *Multimedia Tools and Applications*, vol. 70, no. 1, May 2014, pp. 379–408.
[6] L. Gao *et al.*, "Performance Bound Analysis for Crowdsourced Mobile Video Streaming," *Proc. IEEE CISS*, 2016, pp. 366–71.
[7] M. Tang *et al.*, "A Multi-Dimensional Auction Mechanism For Mobile Crowdsourced Video Streaming," *Proc. IEEE WiOpt*, 2016, pp. 398–405.
[8] M. Tang *et al.*, "MOMD: A Multi-Object Multi-Dimensional Auction for Crowdsourced Mobile Video Streaming," *Proc. IEEE INFOCOM*, 2017.
[9] M. Noura and R. Nordin, "A Survey on Interference Management for Device-To-Device (D2D) Communication and Its Challenges in 5G Networks," *J. Net. Comp. Appl.*, vol. 71, Aug. 2016, pp. 99–117.
[10] C. Zhou, C. W. Lin, and Z. Guo, "mDASH: A Markov Decision-Based Rate Adaptation Approach for Dynamic HTTP Streaming," *IEEE Trans. Multimedia*, vol. 18, no. 4, Apr. 2016, pp. 738–51.
[11] M. J. Neely, "Stochastic Network Optimization with Application to Communication and Queueing Systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, Sept. 2010, pp. 1–211.}
[12] Z. Li et al., "Probe and Adapt: Rate Adaptation for HTTP Video Streaming at Scale," *IEEE JSAC*, vol. 32, no. 4, Mar. 2014, pp. 719–33.
[13] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman, "BOLA: Near-Optimal Bitrate Adaptation For Online Videos," *Proc. IEEE INFOCOM*, 2016, pp. 1–9.
[14] Y. K. Che, "Design Competition through Multidimensional Auctions," *RAND J. Economics*, 1993, pp. 668–80.
[15] M. Zhang *et al.*, "Cooperative and Competitive Operator Pricing for Mobile Crowdsourced Internet Access," *Proc. IEEE INFOCOM*, 2017.

Biographies

Ming Tang [S'16] (tm014@ie.cuhk.edu.hk) is currently pursuing a Ph.D. degree at the Department of Information Engineering, Chinese University of Hong Kong (CUHK). Her research interests include wireless communications and network economics, with current emphasis on user-provided networks, mobile video streaming, and fog computing.

Lin Gao [SM'16] (gaol@hit.edu.cn) is an associate professor in the College of Electronic and Information Engineering at Harbin Institute of Technology, Shenzhen, China. He received his Ph.D. degree in electronic engineering from Shanghai Jiao Tong University in 2010. He won the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2016. His research interests are in the interdisciplinary area of network economics and games.

Haitian Pang [S'16] (pht14@mails.tsinghua.edu.cn) received his B.E. degree in 2014 from the Department of Automation, Tsinghua University, Beijing, China. He is currently pursuing his Ph.D. degree in computer science at Tsinghua University. His research areas include network game modeling, edge content delivery, and multimedia network optimization.

Jianwei Huang [F'16] (jwhuang@ie.cuhk.edu.hk) is an associate professor in the Department of Information Engineering at CUHK. He has received eight international Best Paper Awards, including the IEEE Marconi Prize Paper Award in Wireless Communications 2011. He co-authored six books, including the texbook *Wireless Network Pricing*. He has served as the Chair of IEEE TCCN and MMTC. He is an IEEE ComSoc Distinguished Lecturer and a Thomson Reuters Highly Cited Researcher.

Lifeng Sun [M'05] (sunlf@tsinghua.edu.cn) is a full professor in the Department of Computer Science and Technology, Tsinghua University. His research interests include video streaming, video coding, and multimedia cloud computing. He received the IEEE TCSVT 2010 annul Best Paper Award and the ACM Multimedia 2012 Best Paper Award. He is a member of the IEEE MMC TC and the IEEE VSPC TC.

Frequent disconnections of communications among users can significantly increase the signaling overhead of the cooperation. For example, a helper may find a receiver disconnected after downloading the requested segment. Hence, it is important to design an effective and robust scheduling algorithm by taking into consideration the uncertainty introduced by user mobility.

Fog-Based Transcoding for Crowdsourced Video Livecast

Qiyun He, Cong Zhang, Xiaoqiang Ma, and Jiangchuan Liu

Recent years have witnessed the booming popularity of CLS platforms, through which numerous amateur broadcasters live stream their video contents to viewers around the world. The heterogeneous qualities and formats of the source streams, however, require massive computational resources to transcode them into multiple industrial standard quality versions to serve viewers with distinct configurations, and the delays to the viewers of different locations should be well synchronized to support community interactions.

ABSTRACT

Recent years have witnessed the booming popularity of CLS platforms, through which numerous amateur broadcasters live stream their video contents to viewers around the world. The heterogeneous qualities and formats of the source streams, however, require massive computational resources to transcode them into multiple industrial standard quality versions to serve viewers with distinct configurations, and the delays to the viewers of different locations should be well synchronized to support community interactions. This article attempts to address these challenges and to explore the opportunities with new generation computation paradigms, in particular, fog computing. We present a novel fog-based transcoding framework for CLS platforms to offload the transcoding workload to the network edge (i.e., the massive number of viewers). We evaluate our design through our PlanetLab-based experiment and real-world viewer transcoding experiment.

INTRODUCTION

Recent years have witnessed the emergence of the crowdsourced livecast service (CLS) platforms, represented by Twitch TV (https://www.twitch.tv), YouTube Live (https://www.youtube.com/live), and so on. Given the rapid development of personal computing devices (e.g., smartphones) and broadband network access, most video sources in CLS come from amateur broadcasters rather than commercial/professional content providers, which remarkably stimulates content diversity. This new generation of livecast service has already achieved tremendous success. According to Twitch Retrospective 2015 (https://www.twitch.tv/year/2015), Twitch TV had 35,610 concurrent broadcasters and more than 2 million concurrent viewers during peak time in 2015, with the total stream time exceeding 241 billion minutes through the whole year.

While globally gaming is the major topic on Twitch TV and other mainstream platforms, we have also witnessed the prevalence of regional diversities. For example, the Chinese market is dominated by Douyu TV (https://www.douyu.com), Panda TV (http://www.panda.tv), and Inke (https://www.inke.cn), with a wide range of livecast contents such as singing, talk shows, and even traveling livecast.

Similar to traditional YouTube-like video streaming, the streams in CLS platforms normally have various qualities and formats to serve viewers with diverse network conditions. However, such heterogeneity in source content is more dramatic. For example, as we measured on Twitch TV, the live video contents are generated at over 150 different resolutions, which clearly demands unifying the source streams into industrial standard quality versions. Video on demand (VoD) providers such as YouTube and Netflix have widely adopted adaptive bit rate (ABR) [1] streaming and dynamic adaptive streaming over HTTP (DASH) [2], where the video contents are sliced and transcoded into multiple quality versions, which are finally served to distinct viewers based on their individual requirements. Unlike the VoD scenario, real-time video transcoding for live streaming requires a huge amount of concurrent computational resources, which can be expensive with dedicated servers and even the cloud. As a trade-off between cost and quality of service (QoS), popular CLS platforms such as Twitch TV only offer transcoding services to a small number of premium broadcasters, which make up only 1 to 1.5 percent of all broadcasters.

CLS platforms also involve a great number of viewers who constantly interact with broadcasters and fellow viewers. They have shown great willingness to support the platform and broadcasters, through forms such as donation and monthly subscription at a certain fee. Thanks to the rapid advancement in hardware and software, their home computers nowadays are powerful enough to transcode while playing high-quality live streaming simultaneously. As we observed, the viewer base in major CLS systems is always much larger than the channel base (the number of total channels) at any moment, indicating the existence of a huge amount of potential computational resources.

In this article, we seek novel solutions to offload the massive transcoding workloads in CLS systems to the edge of the network (i.e., the viewers). By analyzing the captured data and systematically studying the user behavior dynamics, we put forward a novel fog-based transcoding framework that crowdsourced the computational resources from viewers and smartly schedule stable viewers for real-time video transcoding. We also address the critical issue of cross-viewer synchronization so as to enhance community interactions.

Digital Object Identifier:
10.1109/MCOM.2017.1600826

The authors are with Simon Fraser University. Xiaoqiang Ma is also with Huazhong University of Science and Technology.

0163-6804/17/$25.00 © 2017 IEEE

Crowdsourced Livecast: Challenges and Opportunities

We first briefly introduce the background and related research on crowdsourced livecast. We then illustrate the unique features, challenges, and opportunities of this new generation of video service through our measurements.

The form of crowdsourced livecast started to get popular in 2012, with some leading platforms guickly expanding in the market. This new generation of user-generated video streaming has also attracted much attention from academia. First, in terms of the social component of these platforms, Kaytoue *et al*. [3] proposed the first characterization of the emerging online Web-based community on Twitch TV, and Hamilton *et al*. [4] further studied its emergence, socialization, and participation. Second, regarding the video streaming performance and user experience, Zhang *et al*. [5] explored the architecture of Twitch TV and investigated the impact of interaction delay on user experience. Aparicio-Pardo *et al*. [6] dug into the Twitch datasets and presented an optimal model for streaming quality to improve viewers' satisfaction. Third, in the field of system architecture, Chen *et al*. [7] proposed a generic cloud renting strategy to optimize the cloud site allocation for video transcoding and delivery in CLS platforms. He et al. [8] further put forward a cloud-based solution that jointly considers user satisfaction and service availability/pricing for video transcoding CLS platforms. However, given the considerable cost of deploying cloud servers and the fact that the CLS platforms charge viewers nothing as a free system by nature, cloud-based transcoding solutions, which are currently adopted by Twitch TV, can only provide transcoding service to very popular streams. To better explore the characteristics and challenges in crowdsourced livecast, we have conducted three measurements.

First, it has been reported that community interaction plays an important role in crowdsourced livecast services. For example, the chat lines of all broadcast channels is found to constantly exceed 400 per second. (http://twitchstatus.com/index.html) Intuitively, it is desirable that fellow viewers are relatively synchronized such that the in-channel community interaction would not cause negative user experience. But according to our measurement, out-of-synchronization chats are not uncommon due to heterogeneous broadcast latency among fellow viewers. We set up two computers: one serves as the broadcaster to encode a source video at 1500 kb/s bit rate; the other hosts two identical virtual machines (VMs). We created a Twitch channel, and the two VMs watched the video stream as viewers. We first set the bandwidth limit to 2000 kb/s for both VMs. We then added the propagation delay of one VM (VM *A*) from 100 to 400 ms, while keeping the other VM (VM *B*) unchanged. The broadcast latency difference between the two viewers under different network conditions is shown in Fig. 1. We can see that, with added propagation delay, the broadcast latency difference becomes significantly larger. A broadcast latency difference of 20 s is almost intolerable for real-time interaction between the viewers. On the other hand, when we changed the bandwidth of VM *A* to

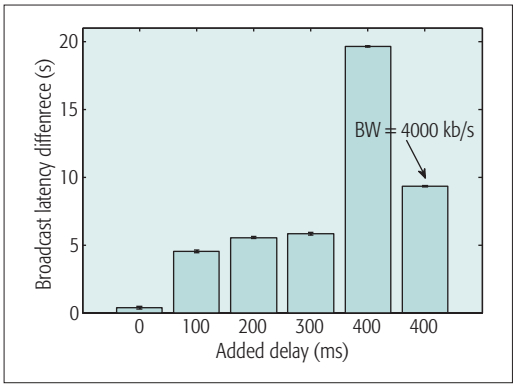


Figure 1. Broadcast delay difference under different network conditions.

4000 kb/s and kept the added propagation delay at 400 ms, the broadcast latency difference is dramatically reduced by half. Our first measurement reveals that divergent end-to-end delays can cause intolerable broadcast latency difference for fellow viewers. Such physical constraints as propagation delay and bandwidth limit, however, are not easy to be lifted for viewers, and we instead should seek for solutions within the broadcast platform.

In our second measurement, we focused on the channels/broadcasters. From February 2015 to June 2015, we captured the data of the broadcasters from Twitch TV every five minutes, using Twitch's public application programming interface (API) (http://dev.twitch.tv). Two outstanding characteristics attracted our attention. First, the resolution of source streams is highly varying over time. For example, at one moment we witnessed 177 different resolutions ranging from 116p to 1600p. Even for the source streams with the same resolution, there are very different bit rates. Second, at any moment the viewer base is dramatically larger than the channel base, and the result is even more remarkable if we only consider the top 10 percent of channels, which attract more than 98 percent of viewers.

The third measurement specifically targets viewer behavior. We captured viewers' online traces of five popular Twitch TV channels from January 25 to February 27, 2015. The measurement captured the JOIN message when any registered viewer joins the channel and the PART message when the viewer leaves the channel. In total, we collected 11,314,442 JOIN records and 11,334,140 PART records. We first see that the overall viewer online duration time can be closely fit to a scaled Pareto distribution function with $\alpha = 0.7$ and $x_m = 2$. Additionally, we can conclude that the longer a viewer is online, the more likely this viewer will continue to be online. The viewers' online duration behavior is also quite consistent, as we see around 80 percent of viewers have a standard deviation less than 20 min.

In summary, in such large CLS platforms as Twitch TV with a massive viewer base, a considerable portion of online viewers are potential resources for stable video transcoding. When exploring these resources, however, inherent challenges lie in the viewers' heterogeneity in terms of their stability and networking/system performance, which calls for effective strategies to distinguish viewers and appropriately make use of their resource. To support rich community inter-

Our first measurement reveals that divergent end-to-end delays can cause intolerable broadcast latency difference for fellow viewers. Such physical constraints as propagation delay and bandwidth limit however are not easy to be lifted for viewers, and we instead should seek for solutions within the broadcast platform.

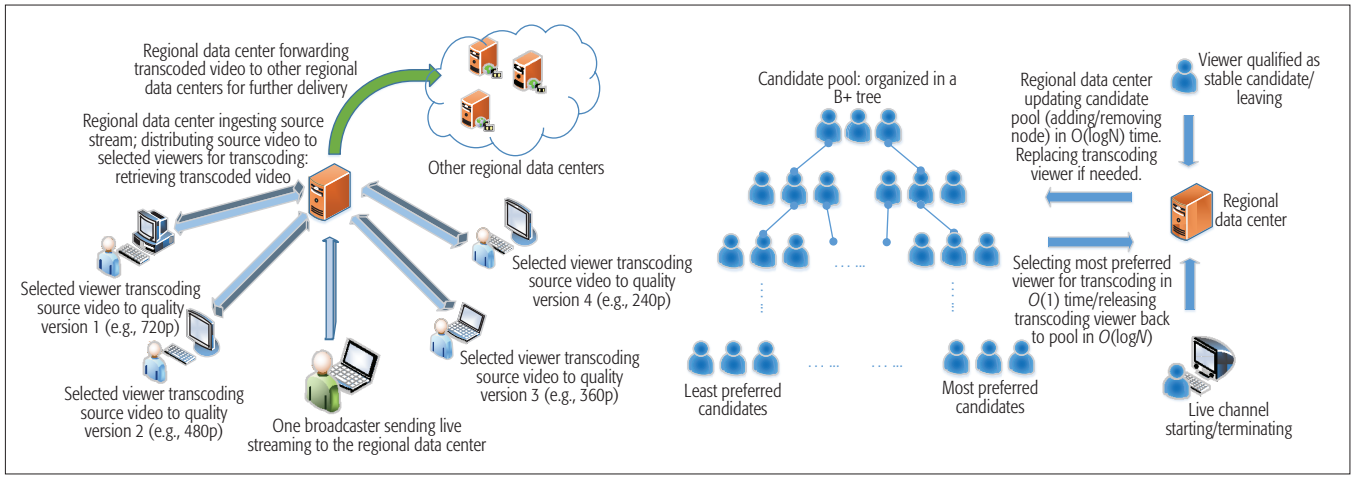


Figure 2. Fog-based transcoding framework: a) system overview; b) scheduling process.

actions, the delays to viewers at different locations should be not only minimized, but also well synchronized. We address these challenges by a novel solution inspired by fog computing [9].

Fog-Based Transcoding: Architectural View

We now illustrate the overall architecture of our design. At the top level, globally the system is divided into multiple regions. Each region is maintained by its own regional data center (also referred to as "regional server"), which is responsible for ingesting source videos from its region, assigning transcoding viewers, recollecting transcoded video, and forwarding the processed streams for further delivery (Fig. 2a). If the regional server cannot find enough transcoding viewers inside the same region, it forwards an assignment request to its neighbor regions, and such assignment is called *cross-region assignment*. If a transcoding viewer becomes offline prior to the termination of the assigned channel, another candidate needs to be scheduled to continue on the transcoding work, which is defined as *reassignment*. At the bottom level, each individual viewer/broadcaster has its own behavior, for example, online/offline time, which is not controlled by the system.

In terms of the scheduling process, our objective is to minimize the number of cross-region assignments and reassignments, as the former introduce extra streaming delay, while the latter cause additional system overhead for re-selecting candidates as well as short absence of the target quality version during reassignment. We therefore want the selected viewers to be as stable as possible, while having a candidate pool of reasonable size so that it will not trigger too many cross-region assignments. Note that we assume selected viewers are cooperative, while in real-world scenarios more incentive mechanisms (e.g., auction for crowdsource [10]) can be used. In the following three subsections, we illustrate how the system selects stable viewers, and then conducts the scheduling process.

Extracting Qualified Stable Candidates

Our observations above indicate that generally the stability of a viewer is proportional to the existing time he/she has already spent in the channel. We therefore set a *waiting threshold* $T(t)$, after passing which the viewer can be regarded as stable. Notably, a longer waiting threshold leads to more stable candidates, but also results in fewer qualified candidates and more time wasted for waiting. We therefore want to maximize the total accumulative transcoding time of stable viewers leaving before and after the channel terminates. To do that, we formulate the mathematical expression of the sum of these two time components and set its derivative to zero, such that the transcoding time is maximized. Given the viewer online duration distribution, the above mentioned scenario is achieved when the waiting threshold is around 30.4 percent of the remaining livecast time. In reality, due to the dynamics of viewers, this result can vary from 25 to 34 percent of the remaining time.

Estimating Individual Stability

Now that we have the optimal online time threshold to filter stable viewers in general, we also expect to estimate the stability of an individual viewer. Some heuristics, such as viewer age, gender, and video quality [11, 12], can be used to further estimate the individual stability in a fine-grained manner. Here we use a simple but effective heuristic that measures the average online duration and standard deviation of a viewer's online record. We use a linear combination of them to further estimate the individual stability. This heuristic particularly reflects the fact that a longer average online duration indicates the viewer tends to stay longer, and a smaller standard deviation means such behavior is more consistent.

Scheduling while Handling Viewer Dynamics

To minimize the reassignment count, we expect to choose the most stable candidate first. However, we are reluctant to change any assignment once made, unless either the directly related viewer or channel state has changed.

To maintain the candidate pool, we can simply re-rank all candidates at any update, but this takes $O(N)$ time every update and $O(1)$ time at assignment time. It incurs significant calculation when N is large with frequent updates. Since there are many more viewer updates than channel updates and reassignments, and re-ranking is only conducted at the latter, we can instead order candidates at the assignment time (taking $O(N$-

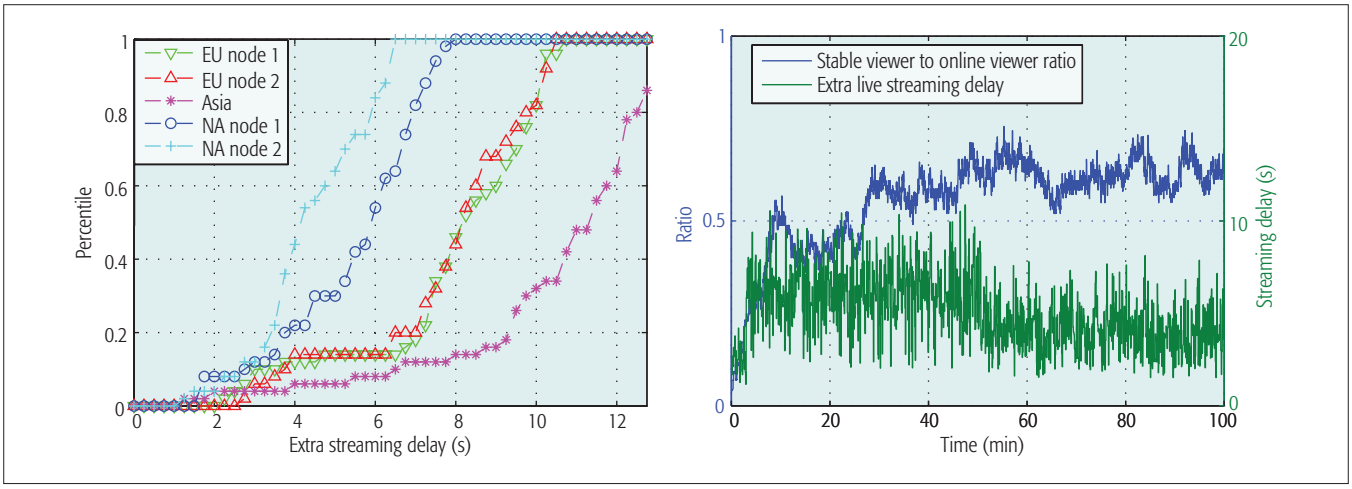


Figure 3. PlanetLab-based experiment result.

log*N*) only). Nonetheless, any extra time at the scheduling moment is undesirable, as it increases the transcoding delay and consequently the live-cast delay. Therefore, we seek to combine both strategies, that is, better organization of the candidates with minimized operational cost per update. Many advanced organization structures can be applied in this context. Using the simple and mature B+ tree as an example, we can deploy a B+ tree in every region to organize all transcoding candidates, where the ordering key is the preference (individual stability) of the candidate. Figure 2b illustrates the major events for its deployment and maintenance, where the the single update operation time is reduced to $O(N\log N)$.

When the system is running, any action will generate a message recording the time, viewer/channel ID, and its type. According to the message, qualified viewers will be inserted into the B+ tree; the starting channel will be assigned with transcoding candidates; the terminating channel will release its transcoding viewers back to the pool; a leaving candidate will be removed from the tree, or be replaced if already assigned.

Prototype and Experimental Study

We implemented a prototype of our fog-based transcoding system on PlanetLab, with 5 nodes as servers, and 208 as viewers and broadcasters. Each viewer node will join the system and stay for a while according to the scaled Pareto distribution. The proposed scheduler will select the most preferred candidates, who then use ffmpeg to transcode a 3.5Mb/s 1080p video into lower-quality versions.

In terms of scheduling results, over time the stable viewer ratio oscillates around 50 to 60 percent, as shown by the blue line in Fig. 3b. We therefore focus more on the streaming performance. We measure the streaming delay, which is one critical metric for video streaming. Figure 3a shows the result measured on the five servers, from which we see huge delay variance in different regions. For example, the delay in North America is much smaller than that in Asia. To further understand the impact of delay variance, we then focus on one server instance in the NA region. As shown in Fig. 3b, the green line indicates the live streaming delay measured by recording the transmission and transcoding time of every 1 s video slice. Clearly, the delay time changes at time 3.5 min and 51 min, which is caused by two reassignment events. The experimental result indicates the existence of delay variance before the transcoded video content is streamed from servers. Although with the optimization of interaction delay the actual delay variance perceived by end viewers is mitigated, it is still worthwhile to make the scheduler take the delay performance into consideration when assigning transcoding workload for the same channel.

To see the transcoding ability of real end viewers instead of PlanetLab virtual machines, we conducted another experiment with eight devices of different popular CPU types. We use *VLC* to do H.264 transcoding for a 1080p video (3.5 Mb/s) to lower-quality versions while the device is playing a live streaming at *Source* quality from a channel on Twitch TV. For comparison, we also measured the transcoding time of the 720p video quality when the device was idle (denoted as 720p*). Table 1 shows the experiment results in the form of transcoding time to video playback time ratio. We see almost all devices can handle quality versions equal to and lower than 480p. Transcoding for the 720p version, however, is more computationally intense, as only the top three devices manage to proceed in real time. Notably, the viewing QoE of all devices is almost not affected, and lack of computational ability is mainly revealed by the long transcoding time. For comparison, we also measured resource usage of a similar workload on an Amazon AWS m3.large server. Typically, transcoding a source video from 1080p to 720p, 480p, 360p, and 240p takes around 73, 54, 42, and 35 percent CPU usage, respectively, which is around the same level of these personal devices.

In short, our experimental result confirms the real-time transcoding ability of modern CPUs, and also suggests distinguishing different unqualified viewers, as some of them can handle lower-quality versions.

Optimizing the Interaction Delay

We now examine the critical issue of cross-viewer synchronization in the CLS platform. We consider a *community* as the broadcaster and the view-

The optimal rate allocation will be obtained in a certain number of iterations. Compared to a centralized solver that directly obtains the optimal solution to the primal problem at the streaming server, our proposed InterSync algorithm is much easier to implement in practical systems.

T CPU type	720p* 2.5 Mb/s	720p 2.5 Mb/s	480p 1.2 Mb/s	360p 0.8 Mb/s	240p 0.5 Mb/s
Intel i7-3770 @3.40 GHz x4	33.7%	59.6%	25.0%	17.5%	14.5%
Intel i7-3630QM @2.4 GHz x4	45.5%	58.2%	33.1%	24.7%	19.7%
Intel i5-2400 3.1 GHz x2	53.5%	66.7%	38.4%	27.8%	19.2%
Intel i5-3210M 2.50 GHz x2	90.6%	113%	68.6%	43.1%	34%
Intel i5-4250U 1.3 GHz x2	116.3%}	191.5%	92%	70.8%	51.2%
AMD a10-4600M 2.3 GHz x2	104.3%	143.3%	77.5%	59.4%	48.2%
Intel i3-2310M 2.10 GHz x2	130.0%	155.8%	90.0%	60.3%	44.4%
Intel Core 2 Duo 2.53 GHz x2	86.7%	190.5%	171.1%	112.3%	76.9%

Table 1. Transcoding time to video playback time ratio of different devices while playing live streaming at Source quality.

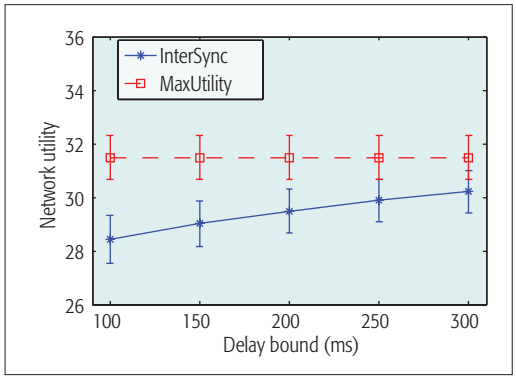


Figure 4. Comparison of network utility.

ers watching the same broadcaster at the same time who are willing to participate in community interaction through chatting. The literal interaction should be relatively synchronized such that viewers' watching experience will not be severely affected. Intuitively, this issue could be solved by dividing viewers into smaller chat channels based on viewers' delay, but this would limit the user interaction in the community. We instead address this issue by adaptively tuning the video rates at viewers to achieve cross-viewer synchronization.

We start from a streaming session consisting of one broadcaster and a set of viewers. A streaming server with certain bandwidth capacity serves the viewers in this session, and the exact video rate allocated to each viewer is adjustable through transcoding, as long as it does not exceed the viewer's downloading bandwidth.

As in previous studies [13, 14], we consider the viewing experience of a viewer is given by a utility function of the allocated video rate, which is strictly concave, increasing, and continuously differentiable. For the streaming session, our objective is then to optimize the viewers' experience through rate adaptation (i.e., tuning streaming rate of each viewer), and meanwhile ensure that the difference between any pair of viewers' network delay is bounded by an empirical threshold. Furthermore, the total streaming rates of all the viewers should not exceed the server's capacity. This leads to a typical network utility maximization (NUM) problem.

Since the objective function is differentiable and strictly concave, and the feasible region is compact, the optimal solution exists (although it may not be unique). This convex problem can be directly solved in a centralized way via the classic simplex and interior-point-based algorithms, provided that the streaming server has the information of each viewer, that is, the end-to-end throughput and the end-to-end delay. It is, however, worth noting that a centralized solver can be very time-consuming for large sessions of thousands of current viewers, not to mention viewers' dynamic join and leave activities.

To this end, we develop a distributed algorithm, *InterSync*, based on dual decomposition, which allows viewers to select appropriate playback bit rates without knowing others' information. The basic idea is that we can solve the dual of the original NUM problem at two levels. At the lower level, each viewer computes its video rate by maximizing its own surplus (i.e., utility minus payment) based on the aggregate price of bandwidth and feeds this value back to the streaming server; at the upper level, the streaming server updates the dual variables according to the feedback information of individual viewers. The optimal rate allocation will be obtained in a certain number of iterations. Compared to a centralized solver that directly obtains the optimal solution to the primal problem at the streaming server, our proposed InterSync algorithm is much easier to implement in practical systems.

We conducted a simulation to compare InterSync with a baseline *MaxUtility*. In MaxUtility, the network utility is maximized subject only to the bandwidth capacity of the streaming server, while the cross-viewer synchronization constraint is not considered. The bandwidth capacity of the streaming server is 100 Mb/s, and the number of viewers is 50. Each viewer's bandwidth is uniformly distributed between 0.5 and 5 Mb/s, and the network delay is uniformly distributed between 50 and 500 ms. The source encoding rate is 5 Mb/s. We vary the end-to-end delay bound δ from 50 to 300 ms.

We report the average and the standard deviation of network utility of both algorithms under different delay bounds, shown in Fig. 4. Since the delay bound is not considered in MaxUtility, the obtained network utility does not change. The network utility obtained by InterSync becomes higher as the delay

bound increases, since the feasible region of the optimization problem (1) also expands with larger delay bound. Although MaxUtility outperforms InterSync in terms of network utility, the real-time community interaction becomes intolerable. In our simulation, the maximum end-to-end delay difference of MaxUtility ranges from 826.0 to 914.4 ms, with an average of 861.4 ms. As evidenced by the measurement above, this level of delay difference would easily lead to tens of seconds broadcast delay among viewers.

Conclusion

In this article, we present a novel fog-based transcoding framework to smartly assign qualified stable viewers to channels for transcoding assignment in CLS systems. We evaluate our design through a PlanetLab-based experiment as well as an end-viewer transcoding experiment. Several future improvements could be studied for our framework. First, it would be interesting to see hybridization of the proposed system with dedicated cloud servers, to have better streaming performance at low cost. Second, more dedicated reward mechanisms can be studied to trigger viewers' incentive of participation. Third, the dynamic pattern inside each region can be studied in a fine-grained manner, after which we can further present strategies for adapting viewer qualifying criteria. Moreover, given that the sources and viewers are shifting to mobile platforms (e.g., Facebook Live and Twitter Periscope for mobile livecast), effectively using their computation resources without significantly increasing their communication costs and energy consumption becomes a challenge too.

Acknowledgment

This publication was made possible by NPRP grant #[8-519-1-108] from the Qatar National Research Fund (a member of Qatar Foundation), and by a Discovery Grant and a Steacie Memorial Fellowship from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] B. Li *et al.*, "Two Decades of Internet Video Streaming: A Retrospective View," *ACM Trans. Multimedia Computing Commun. Applications*, vol. 9, no. 1s, article no. 33, Oct. 2013.

[2] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate Adaptation for Adaptive HTTP Streaming," *Proc. ACM MMSys 2011*.

[3] M. Kaytoue *et al.*, "Watch Me Playing, I Am a Professional: A First Study on Video Game Live Streaming," *Proc. ACM WWW 2012*.

[4] W. A. Hamilton, O. Garretson, and A. Kerne, "Streaming on Twitch: Fostering Participatory Communities of Play within Live Mixed Media," *Proc. ACM CHI 2014*.

[5] C. Zhang and J. Liu, "On Crowdsourced Interactive Live Streaming: A Twitch TV-Based Measurement Study," *Proc. ACM NOSSDAV 2015*.

[6] R. Aparicio-Pardo *et al.*, "Transcoding Live Adaptive Video Streams at a Massive Scale on the Cloud," *Proc. ACM MMSys 2015*.

[7] F. Chen *et al.*, "Crowdsourced Live Streaming over the Cloud," *Proc. IEEE INFOCOM 2015*.

[8] Q. He *et al.*, "Coping with Heterogeneous Video Contributors and Viewers in Crowdsourced Live Streaming: A Cloud-Based Approach," *IEEE Trans. Multimedia*, vol. 18, no. 5, May 2016, pp. 916–28.

[9] F. Bonomi *et al.*, "Fog Computing and Its Role in the Internet of Things," *Proc. ACM MCC 2012*.

[10] Y. Singer and M. Mittal, "Pricing Mechanisms for Crowdsourcing Markets," *Proc. ACM WWW 2013*.

[11] F. Dobrian *et al.*, "Understanding the Impact of Video Quality on User Engagement," *Proc. ACM SIGCOMM Comp. Commun. Review*, vol. 41, no. 4, Aug. 2011, pp. 362–73.

[12] S. S. Krishnan and R. K. Sitaraman, "Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-Experimental Designs," *IEEE/ACM Trans. Networking*, vol. 21, no. 6, Feb. 2013, pp. 2001–14.

[13] J. Huang *et al.*, "Joint Source Adaptation and Resource Allocation for Multi-User Wireless Video Streaming," *IEEE Trans. Circuits Systems Video Technology*, vol. 18, no. 5, May 2008, pp. 582–95.

[14] M. Chen *et al.*, "Utility Maximization in Peer-To-Peer Systems," *ACM SIGMETRICS Performance Evaluation Review*, vol. 36, no. 1, Dec. 2008, pp. 169–80.

Biographies

Qiyun He (qiyunh@cs.sfu.ca) received a B.Eng. degree from Zhejiang University, China, and a B.Sc. degree from Simon Fraser University, British Columbia, Canada, both in 2015. He is currently an M.Sc. student at the School of Computing Science, Simon Fraser University. His research interests include cloud computing, crowdsourcing, multimedia systems, and networks.

Cong Zhang (congz@cs.sfu.ca) received his M.S. degree in information engineering from Zhengzhou University, China, in 2012, and is currently working toward a Ph.D. degree in computing science at Simon Fraser University. He is currently working with the Network Modeling Research Group, Simon Fraser University. His research interests include multimedia communications, cloud computing, and crowdsourced live streaming.

Xiaoqiang Ma (xma10@cs.sfu.ca) received his B.Eng. degree from Huazhong University of Science and Technology, Wuhan, China, in 2010, and his M.Sc. and Ph.D. degrees from Simon Fraser University in 2012 and 2015, respectively. He is currently an assistant professor with the School of Electronic Information and Communication, Huazhong University of Science and Technology. His research interests include wireless networking, multimedia, cloud, and big data.

Jiangchuan Liu [F] (jcliu@cs.sfu.ca) is currently a full professor (with a University Professorship) in the School of Computing Science at Simon Fraser University. He is an NSERC E.W.R. Steacie Memorial Fellow. He is a Steering Committee Member of *IEEE Transactions on Mobile Computing*, and an Associate Editor of *IEEE/ACM Transactions on Networking*, *IEEE Transactions on Big Data*, and *IEEE Transactions on Multimedia*.

Given that the sources and viewers are shifting to mobile platforms (e.g., Facebook Live and Twitter Periscope for mobile livecast), effectively using their computation resources without significantly increasing their communication costs and energy consumption becomes a challenge too.

Coding for Distributed Fog Computing

Songze Li, Mohammad Ali Maddah-Ali, and A. Salman Avestimehr

Redundancy is abundant in fog networks (i.e., many computing and storage points) and grows linearly with network size. The authors demonstrate the transformational role of coding in fog computing for leveraging such redundancy to substantially reduce the bandwidth consumption and latency of computing.

ABSTRACT

Redundancy is abundant in fog networks (i.e., many computing and storage points) and grows linearly with network size. We demonstrate the transformational role of coding in fog computing for leveraging such redundancy to substantially reduce the bandwidth consumption and latency of computing. In particular, we discuss two recently proposed coding concepts, minimum bandwidth codes and minimum latency codes, and illustrate their impacts on fog computing. We also review a unified coding framework that includes the above two coding techniques as special cases, and enables a trade-off between computation latency and communication load to optimize system performance. At the end, we will discuss several open problems and future research directions.

INTRODUCTION

The fog architecture (Fig. 1) has been proposed recently to better satisfy the service requirements of the emerging Internet of Things (IoT) (e.g., [1]). Unlike cloud computing, which stores and processes end users' data in remote and centralized data centers, fog computing brings the provision of services closer to the end users by pooling the available resources at the edge of the network (e.g., smartphones, tablets, smart cars, base stations, and routers) (e.g., [2, 3]). As a result, the main driving vision for fog computing is to leverage the significant amount of dispersed computing resources at the edge of the network to provide much more user-aware, resource-efficient, scalable, and low-latency services for IoT.

The main goal of this article is to demonstrate how coding can be effectively utilized to trade abundant computing resources at the network edge for communication bandwidth and latency. In particular, we illustrate two recently proposed novel coding concepts that leverage the available or underutilized computing resources at various parts of the network to enable coding opportunities that significantly reduce the bandwidth consumption and latency of computing, which are of particular importance in fog computing applications.

The first coding concept, introduced in [4, 5], which we refer to as *minimum bandwidth codes*, enables a surprising inverse-linear trade-off between computation load and communication load in distributed computing. Minimum bandwidth codes demonstrate that increasing the computation load by a factor of r (i.e., evaluating each computation at r carefully chosen nodes) can create novel coding opportunities that reduce the required communication load for computing by the same factor. Hence, minimum bandwidth codes can be utilized to pool the underutilized computing resources at the network edge to slash the communication load of fog computing.

The second coding concept, introduced in [6], which we refer to as *minimum latency codes*, enables an inverse-linear trade-off between computation load and computation latency (i.e., the overall job response time). More specifically, minimum latency codes utilize coding to effectively inject redundant computations to alleviate the effects of stragglers and speed up the computations by a multiplicative factor that is proportional to the amount of injected redundancy. Hence, by utilizing more computation resources at the network edge, minimum latency codes can significantly speed up distributed fog computing applications.

In this article, we give an overview of these two coding concepts, illustrate their key ideas via motivating examples, and demonstrate their impacts on fog networks. More generally, noting that redundancy is abundant in fog networks (i.e., many computing/storage points) and grows linearly with network size, we demonstrate the transformational role of coding in fog computing for leveraging such redundancy to substantially reduce the bandwidth consumption and latency of computing. We also point out that while these two coding techniques are also applicable to cloud computing applications, they are expected to play a much more substantial role in improving the system performance of fog applications, due to the fact that the communication bottleneck and straggling nodes are far more severe issues in fog computing compared to its cloud counterpart.

We also discuss a recently proposed *unified* coding framework, in which the above two coding concepts are systematically combined by introducing a trade-off between computation latency and communication load. This framework allows a fog computing system to operate at any point on the trade-off, on which the minimum bandwidth codes and minimum latency codes can be viewed as two extreme points that minimizes the communication load and computation latency, respectively.

We finally conclude the article and highlight some exciting open problems and research directions for utilizing coding in fog computing architectures.

Digital Object Identifier:
10.1109/MCOM.2017.1600894

Songze Li and A. Salman Avestimehr are with the University of Southern California; Mohammad Ali Maddah-Ali is with Nokia Bell Labs.

Minimum Bandwidth Codes

We illustrate minimum bandwidth codes in a typical fog computing scenario, in which a fog client aims to utilize the network edge for its computation task. For instance, a driver wants to find the best route through a navigation application offered by the fog, in which the map information and traffic condition are distributedly stored in edge nodes (ENs) like roadside monitors, smart traffic lights, and smart cars that collaborate to find the best route. Another example is object recognition, which is the key enabler of many augmented reality applications. To provide an object recognition service over the fog, edge nodes like routers and base stations each stores parts of the dataset repository, and they collaboratively process the images or videos provided by the fog client.

For the above fog computing applications, the computation task is over a large dataset that is distributedly stored on the ENs (e.g., map/traffic information or dataset repository), and the computations are often decomposed using MapReduce-type frameworks (e.g., [7, 8]), in which a collection of ENs distributedly *map* a set of input files, generating some intermediate values, from which they *reduce* a set of output functions.

We now demonstrate the main concepts of minimum bandwidth codes in a simple problem depicted in Fig. 2. In this case, a client uploads a job of computing three output functions (represented by red/circle, green/square, and blue/triangle) from six input files to the fog. Three edge nodes in the fog (EN 1, EN 2, and EN 3) collaborate to perform the computation. Each EN is responsible for computing a unique output function, for example, EN 1 computes the red/circle function, EN 2 computes the green/square function, and EN 3 computes the blue/triangle function. When an EN maps a locally stored input file, it computes three intermediate values, one for each output function. To reduce an output function, each EN needs to know the intermediate values of this output for all six input files.

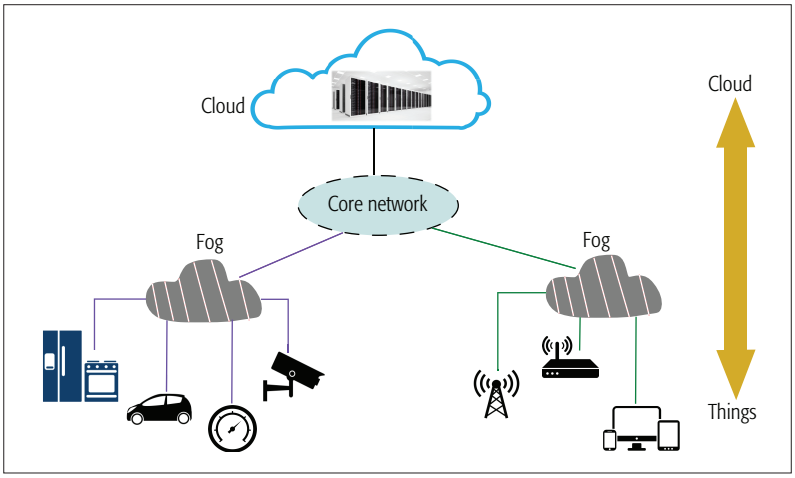


Figure 1. Illustration of a fog architecture.

We first consider the case where no redundancy is imposed on the computations, that is, each file is mapped exactly once. Then, as shown in Fig. 2a, each EN maps two input files locally, obtaining two out of six required intermediate values. Hence, each EN needs another four intermediate values transferred from the other ENs, yielding a communication load of $4 \times 3 = 12$.

Now, we demonstrate how minimum bandwidth codes can substantially reduce the communication load by injecting redundancy in computation. As shown in Fig. 2b, let us double the computation such that each file is mapped on two ENs (files are downloaded to the ENs offline). It is apparent that since more local computations are performed, each EN now only requires two other intermediate values, and an uncoded shuffling scheme would achieve a communication load of $2 \times 3 = 6$. However, we can do better with the minimum bandwidth codes. As shown in

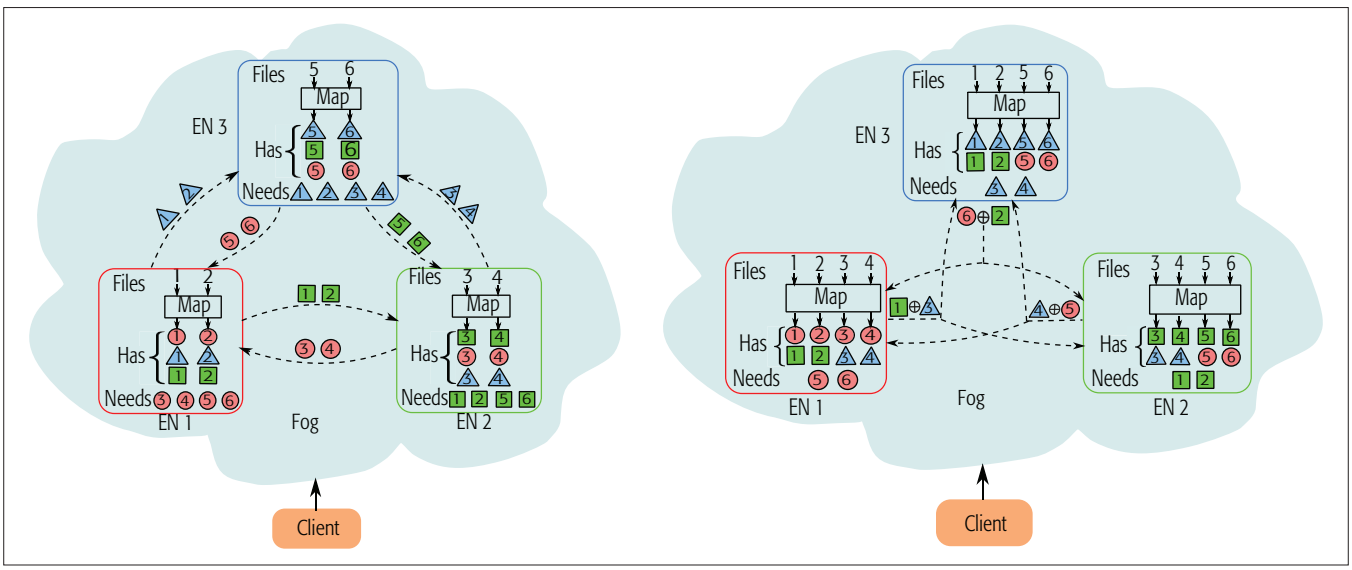


Figure 2. a) An uncoded fog computing scheme to compute three functions, one on each of the three ENs, from six files. Each file is mapped once on one EN, and each EN has four intermediate values transferred uncodedly from the other ENs to reduce the corresponding output; b) implementation of a minimum bandwidth code on three ENs. Each of the six files is mapped on two ENs. During data shuffling, each EN creates a coded packet that is simultaneously useful for the other two ENs by XORing two locally computed intermediate values, and multicasts the packet to the other two ENs.

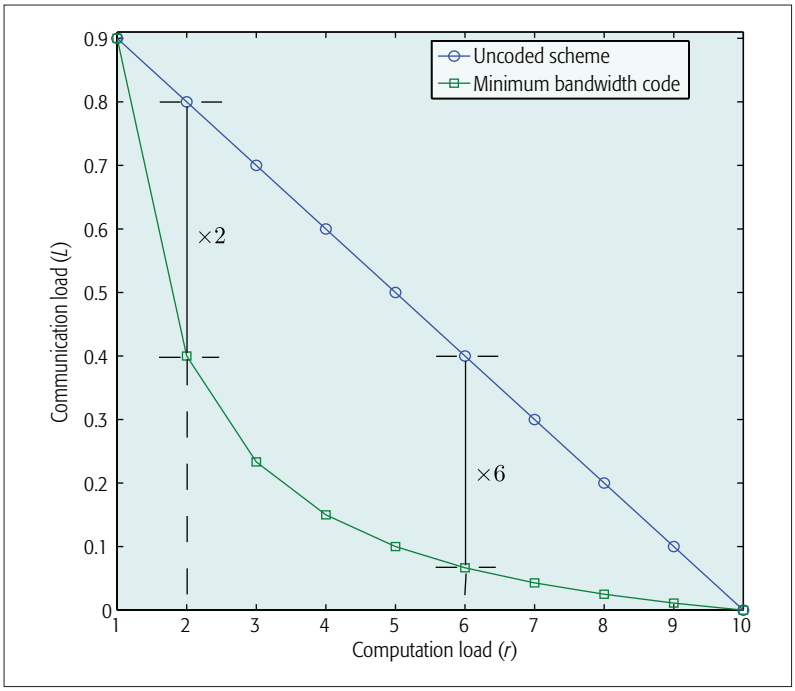


Figure 3. Comparison of the communication load of minimum bandwidth codes with that of the uncoded scheme, for a network with $K = 10$ ENs.

Fig. 2b, instead of unicasting individual intermediate values, every EN multicasts a bit-wise XOR, denoted by $\oplus$, of two intermediate values to the other two ENs, simultaneously satisfying their data demands. For example, knowing the blue triangle in file 3, EN 2 can cancel it from the coded packet multicast by EN 1, recovering the needed green square in file 1. In general, the bandwidth consumption of multicasting one packet to two nodes is less than that of unicasting two packets, and here we consider a scenario in which it is as much as that of unicasting one packet (which is the case for wireless networks). Therefore, the above minimum bandwidth code incurs a communication load of 3, achieving a 4× gain from the case without computation redundancy and a 2× gain from the uncoded shuffling.

More generally, we can consider a fog computing scenario in which K ENs collaborate to compute Q output functions from N input files that are distributedly stored at the nodes. We define the computation load, r, to be the total number of input files that are mapped across the nodes, normalized by N. For example, $r = 2$ means that on average each file is mapped on two nodes. We can similarly define the communication load L to be the total (normalized) number of information bits exchanged across nodes during data shuffling in order to compute the Q output functions.

For this scenario it was shown in [5] that, compared to conventional uncoded strategies, minimum bandwidth codes can surprisingly reduce the communication load by a multiplicative factor that equals the computation load r when computing r times more sub-tasks than the execution without redundancy (i.e., $r = 1$). Or more specifically,

$$L_{\text{coded}} = \frac{1}{r} L_{\text{uncoded}} = \frac{1}{r}\left(1 - \frac{r}{K}\right) = \Theta\left(\frac{1}{r}\right). \quad (1)$$

Minimum bandwidth codes employ a specific strategy to assign the computations of the map and reduce functions in order to enable novel coding opportunities for data shuffling. In particular, each data block is repetitively mapped on r distinct nodes according to a specific pattern in order to create coded multicast messages that deliver useful data simultaneously to $r \geq 1$ nodes. For example, as demonstrated in Fig. 3, the overall communication load can be reduced by more than 50 percent when each map task is repeated at only one other node (i.e., $r = 2$).

The idea of efficiently creating and exploiting coded multicast opportunities was initially proposed to solve caching problems in [9, 10], and extended to wireless device-to-device (D2D) networks in [11], where caches pre-fetch part of the content to enable coding during the content delivery, minimizing the network traffic. Minimum bandwidth codes extend such coding opportunities to data shuffling of distributed computing frameworks, significantly reducing the required communication load.

Apart from significantly slashing the bandwidth consumption, minimum bandwidth codes also have the following major impacts on the design of fog computing architecture.

Reducing Overall Response Time

Let us consider an arbitrary fog computing application for which the overall response time is composed of the time spent computing the intermediate tasks, denoted by $T_{\text{Task Computation}}$, and the time spent moving intermediate results, denoted by $T_{\text{Data Movement}}$. In many applications of interest (e.g., video/image analytics or recommendation services), most of the job execution time is spent on data movement. For example, consider the scenarios in which $T_{\text{Data Movement}}$ is $10\times \sim 100\times$ $T_{\text{Task Computation}}$. Using a minimum bandwidth code with computation load r, we can achieve an overall response time of

$$T_{\text{total, coded}} \approx \mathbb{E}\left[rT_{\text{Task Computation}} + \frac{1}{r} T_{\text{Data Movement}} \right]. \quad (2)$$

To minimize the above response time, one would choose the optimum computation load,

$$r^* = \sqrt{\frac{T_{\text{Data Movement}}}{T_{\text{Task Computation}}}}.$$

Then in the above example, utilizing minimum bandwidth codes can reduce the overall job response time by approximately $1.5 \sim 5$ times.

The impact of minimum bandwidth codes on reducing the response time has been demonstrated recently in [12] through a series of experiments over Amazon EC2 clusters. In particular, the minimum bandwidth codes were incorporated into the well-known distributed sorting algorithm `TeraSort` [13] to develop a new coded sorting algorithm, `CodedTeraSort`, which allows a flexible selection of the computation load r. In Table 1, we summarize the runtime performance of a particular job of sorting 12 GB of data over 16 EC2 instances.

Theoretically according to Eq. 1, with a computation load $r = 5$, `CodedTeraSort` promises to reduce the data shuffling time by a factor of approximately 5. From Table 1, we can see that

	CodeGen (sec.)	Map (sec.)	Pack/Encode (sec.)	Shuffle (sec.)	Unpack/ Decode (sec.)	Reduce (sec.)	Total Time (sec.)	Speedup
`TeraSort:`	–	1.86	2.35	945.72	0.85	10.47	961.25	
`CodedTeraSort: r = 5`	23.47	10.84	8.10	222.83	3.69	14.40	283.33	3.39

Table 1. Average response times for sorting 12 GB of data over 16 EC2 instances using 100 Mb/s network speed.

while computing $r = 5$ times more map functions increased the map task computation time by 5.83×, `CodedTeraSort` brought down the data shuffling time, which was the limiting component of the runtime of this application, by 4.24×. As a result, `CodedTeraSort` reduced the overall job response time by 3.39×.

Scalable Mobile Computation

The minimum bandwidth codes also found application in a wireless distributed computing platform proposed in [14], which is a fully decentralized fog computing environment. In this platform, a collection of mobile users, each with input to process overall a large dataset (e.g., the image repository of an image recognition application), collaborate to store the dataset and perform the computations using their own storage and computing resources. All participating users communicate the locally computed intermediate results among each other to reduce the final outputs.

Utilizing minimum bandwidth codes in this wireless computing platform leads to a *scalable* design. More specifically, let us consider a scenario where K users, each processing a fraction of the dataset, denoted by μ (for some $(1/K) \leq \mu \leq 1$), collaborate for wireless distributed computing. It is demonstrated in [14] that minimum bandwidth codes can achieve a (normalized) bandwidth consumption of $(1/\mu) - 1$ to shuffle all required intermediate results. This reduces the communication load of the uncoded scheme, that is, $K(1 - \mu)$, by a factor of μK, which scales linearly with the aggregated storage size of all collaborating users. Also, since the consumed bandwidth is independent of the number of users K, minimum bandwidth code allows this platform to simultaneously serve an unlimited number of users with a constant communication load.

Minimum Latency Codes

We now move to the second coding concept, minimum latency codes, and demonstrate it for a class of fog computing applications in which a client's input is processed over a large dataset (possibly over multiple iterations). The application is supported by a group of ENs, which have distributedly stored the entire dataset. Each node processes the client's input using the parts of the dataset it locally has, and returns the computed results to the client. The client reduces the final results after collecting intermediate results from all ENs. Many distributed machine learning algorithms fall into this category. For example, a gradient descent algorithm for linear regression requires multiplying the weight vector with the data matrix in each iteration. To do that at the network edge, each EN locally stores a sub-matrix of the data matrix. During computation, each EN multiplies the weight vector with the stored sub-matrix and returns the results to the client.

To be more specific, let us consider a simple distributed matrix multiplication problem in which, as shown in Fig. 4, a client wants to multiply a data matrix $\mathbf{A}$ with the input matrix $\mathbf{X}$ to compute $\mathbf{AX}$. The data matrix $\mathbf{A}$ is stored distributedly across three nearby ENs (EN 1, EN 2, and EN 3), on which the matrix multiplication will be executed distributedly.

One natural approach to tackle this problem is to vertically and evenly divide data matrix $\mathbf{A}$ into three sub-matrices, each of which is stored on one EN. Then when each EN receives input $\mathbf{X}$, it simply multiplies its locally stored sub-matrix by $\mathbf{X}$ and returns the results, and the client vertically concatenates the returned matrices to obtain the final result. However, we note that since this uncoded approach relies on successfully retrieving the task results from all three ENs, it has a major drawback that once one of the ENs runs slow or gets disconnected, the computation may take very long or even fail to finish. Minimum latency codes deal with slow or unreliable ENs by optimally creating redundant computation tasks. As shown in Fig. 4, a minimum latency code vertically partitions data matrix $\mathbf{A}$ into two sub-matrices, $\mathbf{A}_1$ and $\mathbf{A}_2$, and creates one redundant task by summing $\mathbf{A}_1$ and $\mathbf{A}_2$. Then $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_1 + \mathbf{A}_2$ are stored on EN 1, EN 2, and EN 3, respectively. In the case of Fig. 4, the computation is completed when the client has received the task results only from ENs 1 and 3, from which $\mathbf{A}_2\mathbf{X}$ can be decoded. In fact, it is obvious that the client can recover the final result once she receives the task results from any two of the three ENs, without needing to wait for the slow/unreachable EN (EN 2 in this case). In summary, minimum latency codes create redundant computation tasks across fog networks, such that having *any* set of a certain number of task results is sufficient to accomplish the overall computation. Hence, applying minimum latency codes on the abundant ENs can effectively alleviate the effect of stragglers and significantly speed up fog computing.

As illustrated in the above example, the basic idea of minimum latency codes is to apply erasure codes on computation tasks, creating redundant coded tasks that provide robustness to straggling ENs. Erasure codes have been widely exploited to combat symbol losses in communication systems and disk failures in distributed storage systems. The simplest form of erasure code, the repetition code, repeats each information symbol multiple times, such that a information symbol can be successfully recovered as long as at least one of the repeats survives. For example, modern distributed file systems like the Hadoop distributed file system (HDFS) replicate each data block three times across different storage nodes. Another type of erasure code, known as the maximum-distance-separable (MDS) code, provides better robustness to erasures. An (n, k) MDS code takes k information symbols and encodes them into $n \geq k$ coded symbols, such that obtaining *any* k out

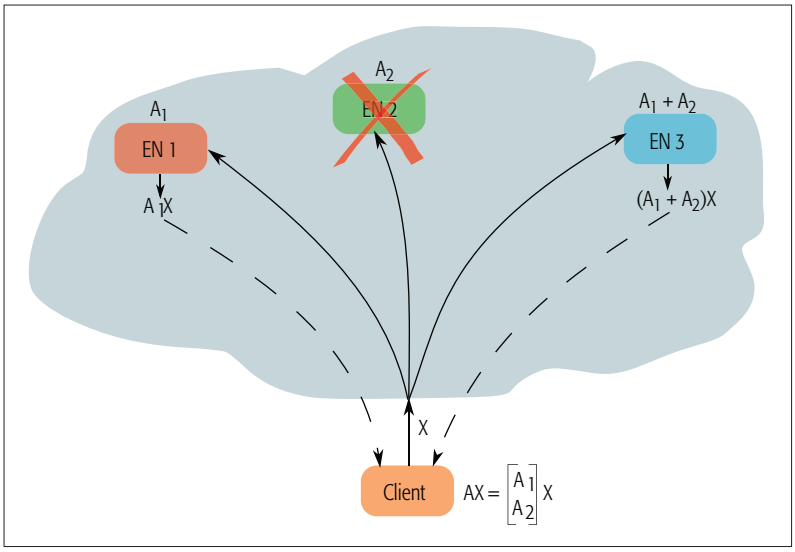


Figure 4. Matrix multiplication using a minimum latency code. The code generates three coded tasks, each executed by an EN. The client recovers the final result after receiving the results from EN 1 and EN3. The runtime is not affected by the straggling EN 2.

of the n coded symbols is sufficient to decode all k information symbols. This "any k of n" property is highly desirable due to the randomness of erasures. A successful application of the MDS code is the Reed-Solomon code used to protect CDs and DVDs.

As introduced in [6], minimum latency codes are exactly MDS codes that are used to encode computation tasks. For a fog computing job executed on n ENs, an (n, k) minimum latency code first decomposes the overall computation into k smaller tasks, for some $k \leq n$. Then it encodes them into n coded tasks using an (n, k) MDS code, and assigns each of them to a node to compute. By the aforementioned "any k of n" property of the MDS code, we can accomplish the overall computation once we have collected the results from the *fastest* k out of n coded tasks, without worrying about the tasks still running on the slow nodes (or stragglers).

Minimum latency codes can help to significantly improve the response time of fog applications. Let's consider a computation task performed distributedly across n ENs. The response time of the uncoded approach is limited by the slowest node. An (n, k) repetition code breaks the computation into k tasks, and repeats each task (n/k) times across the n nodes, and the computation continues until each task has been computed at least once. On the other hand, for an (n, k) minimum latency code, the response time is limited by the fastest k out of n nodes that have finished their coded tasks. As shown in [6], for a shifted-exponential distribution, the average response times of the uncoded execution and the repetition code are both

$$\Theta\left(\frac{\log n}{n}\right).$$

The minimum latency codes can reduce the response time by a factor of $\Theta(\log n)$. For example, in a typical fog computing scenario with 10 $\sim$ 100 nodes, minimum latency codes can theoretically offer a 2.3× $\sim$ 4.6× speedup. Moreover, experiments on Amazon EC2 clusters were performed in [6], in which for a gradient descent computation for linear regression, minimum latency codes reduce the response time by 35.7 percent on average. We further envision that in a fog computing environment where computing nodes are much more heterogeneous and likely to be unresponsive, the performance gain by using minimum latency codes will be much larger.

Other than speeding up the fog computing applications, minimum latency codes also maximize the survivability of the computation when faced with nodes failure/disconnection, that is, when the task results may never come back. We note that an (n, k) minimum latency code requires any k out of n tasks to be returned to guarantee a successful computation, and this level of robustness cannot be provided by either the uncoded computation or the repetition code.

A Unified Coding Framework

We have so far discussed two different coding techniques that aim at minimizing the bandwidth consumption and the computation latency of fog computing. However, under a MapReduce-type computing model, a *unified* coded framework has been developed recently in [15] by introducing a trade-off between computation latency in the map phase and communication load in the shuffle phase.

As an example, in Fig. 5 we illustrate the trade-off between computation latency and communication load that is achieved by the unified framework for running a distributed matrix multiplication over 18 edge nodes (see [15, Sec. III] for details). We observe that the achieved trade-off approximately exhibits an inverse-linearly proportional relationship between the latency and the load. In particular, we can see that the minimum bandwidth codes and minimum latency codes can be viewed as special instances of the proposed coding framework by considering two extremes of this trade-off: minimizing either the communication load or the computation latency individually. Next, we further illustrate how to utilize this trade-off to minimize the total response time, which is the sum of the communication time in the shuffle phase and the computation latency in the map phase. For the matrix multiplication problem in Fig. 5, we consider real entries, each represented using 2 bytes, a shift-exponential distribution for the map task execution time, and a wireless network with speed 10 Mb/s. Then the minimum bandwidth codes that wait for all 18 nodes to finish their map tasks achieve a total response time of 302 s,[1] and the minimum latency codes that terminate the map phase when the fastest 3 nodes (minimum required number) finish their map tasks achieve a total response time of 263 s. Using the unified coding framework, we can wait for the optimal number of the fastest 12 nodes to finish, and achieve the minimum total response time of 186 s. Hence, this unified coding approach provides a performance gain of 38.4 and 29.3 percent over the minimum bandwidth codes and minimum latency codes respectively.

This unified coding framework, which is essentially a systematic concatenation of the minimum bandwidth codes and minimum latency codes, takes advantage of both coding techniques in

[1] The communication load in Fig. 5 is normalized by the number of the rows of the matrix, which is 10^6 in this example.

different stages of the computation. In the map phase, MDS codes are employed to create coded tasks, which are then assigned to ENs in a specific repetitive pattern for local execution. According to the specific computation latency of the map phase, all running map tasks are terminated as soon as a certain number of nodes have finished their local computations. Then in the shuffle phase, coded multicast opportunities specified by minimum bandwidth codes are greedily utilized until the data demands of all nodes are satisfied. For example, we can consider executing a linear computation consisting of $m = 20$ map tasks using $K = 6$ edge nodes, each of which can process $\mu = (1/2)$ fractions of the tasks. To be able to end the map phase when only the fastest $q = 4$ nodes finish their local tasks, we can first use a $((K/q)m, m) = (30, 20)$ MDS code to generate 30 coded tasks, each of which is then assigned to $\mu q = 2$ nodes for execution according to the repetitive assignment pattern specified by the minimum bandwidth codes. For more detailed illustrative examples, we refer the interested readers to [15, Sec. IV].

The unified coding framework allows us to flexibly select the operation point to minimize the overall job execution time. For example, when the network is slow, we can wait for more nodes to finish their map computations, creating better multicast opportunities to further slash the amount of data movement. On the other hand, when we have detected that some nodes are running slow or becoming unresponsive, we can shift the load to the network by ending the map phase as soon as enough coded tasks are executed.

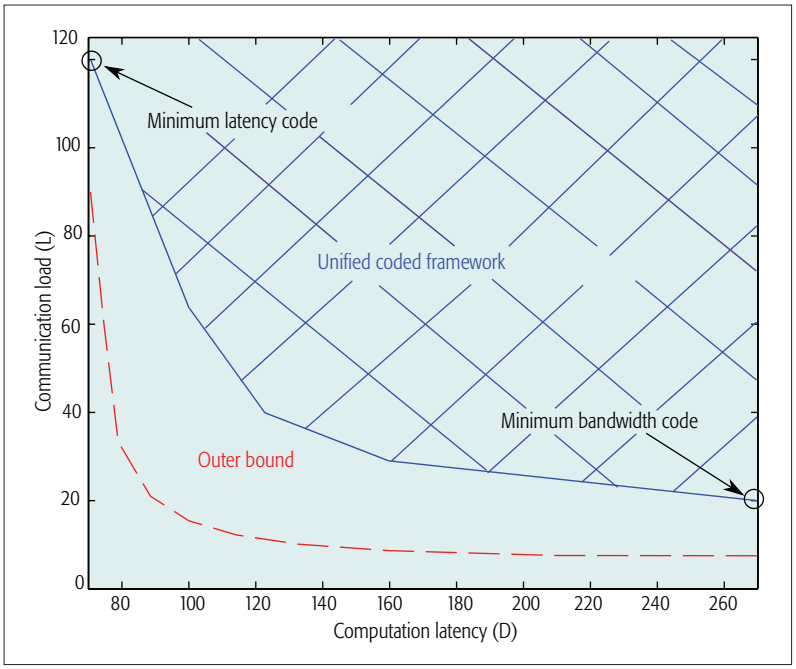


Figure 5. Comparison of the latency-load pairs achieved by the proposed unified scheme with the outer bound, for a distributed matrix multiplication job over a network of 18 nodes.

Conclusions and Future Research Directions

We demonstrate how coding can be effectively utilized to leverage abundant computing resources at the network edge to significantly reduce the bandwidth consumption and computation latency in fog computing applications. In particular, we illustrate two recently proposed coding concepts, minimum bandwidth codes and minimum latency codes, and discuss their impacts on fog computing. We also discuss a unified coding framework that includes the above two coding techniques as special cases, and enables a trade-off between computation latency and communication load to optimize the system performance.

We envision codes playing a fundamental role in fog computing by enabling efficient utilization of computation, communication, and storage resources at the network edge. This area opens up many important and exciting future research directions. Here we list a few.

Heterogeneous Computing Nodes

In distributed fog networks, different nodes have different processing and storage capacities. The ideas outlined in this article can be used to develop heuristic solutions for heterogeneous networks. For example, one simple approach is to break the more powerful nodes into multiple smaller virtual nodes that have homogeneous capability, and then apply the proposed coding techniques to the homogeneous setting. However, systematically developing practical task assignment and coding techniques for these systems that are provably optimum (approximately) is a challenging open problem.

Networks with Multi-Layer and Structured Topology

The current code designs for distributed computing[4, 5, 15] are developed for a basic topology in which the processing nodes are connected through a shared link. While these results demonstrate the significant gain of coding in distributed fog computing, we need to extend these ideas to more general network topologies. In such networks, nodes can be connected through multiple switches and links in different layers with different capacities.

Multi-Stage Computation Tasks

Another important direction is to consider more general computing frameworks, in which the computation job is represented by a directed acyclic task graph (DAG). While we can apply the aforementioned code designs for each stage of computation locally, we expect to achieve a higher reduction in bandwidth consumption and response time by globally designing codes for the entire task graph and accounting for interactions between consecutive stages.

Coded Computing Overhead

The current fog computing system under consideration lacks appropriate modeling of the coding overhead, which includes the cost of the encoding and decoding processes, the cost of performing multicast communications, and the cost of maintaining desired data redundancy across fog nodes. To make the study of coding in practical fog systems more relevant, it is important to carefully formulate a comprehensive model that systematically accounts for these overheads.

Fog architecture facilitates offloading of computational tasks from relatively weak computational devices (clients) to more powerful nodes in the edge network. As a result, there is a critical need for "verifiable computing" methods, in which clients can make sure they receive the correct calculations.

Verifiable Distributed Computing

Fog architecture facilitates offloading of computational tasks from relatively weak computational devices (clients) to more powerful nodes in the edge network. As a result, there is a critical need for "verifiable computing" methods, in which clients can make sure they receive the correct calculations. This is typically achieved by injecting redundancy in computations by the clients. We expect codes to provide much more efficient methods for leveraging computation redundancy in order to provide verified computing in fog applications.

Exploiting the Algebraic Structures of Computation Tasks

Recall that the minimum bandwidth codes can be applied to any general computation task that can be cast in a MapReduce framework. However, we expect to improve the overall performance if we exploit the specific algebraic properties of the underlying tasks. For example, if the task has some linearity, we may be able to incorporate it in communication and coding design in order to further reduce the bandwidth consumption and latency. On the contrary, minimum latency codes work only for some particular linear functions (e.g., matrix multiplication). It is of great interest to extend these codes to a broader class of computation tasks.

Communication-Heavy Applications

Recall that by exploiting minimum bandwidth codes we can envision a fog system that can handle many distributed fog nodes with a bounded communication load. Such a surprising feature would enormously expand the list of applications that can be offered over fog networks. One research direction is to re-examine some communication-heavy tasks to see if minimum bandwidth codes allow them to be implemented over distributed fog networks.

Plug-and-Play Fog Nodes

We can finally envision a software package (or app) that can be installed and maintained distributedly on each fog node. This package should allow a fog computing node to join the system anytime to work with the rest of the nodes or leave the system asynchronously, while the entire network still operates near optimum. Designing codes that guarantee integrity of computations despite such network dynamics is a very interesting and important research direction.

References

[1] F. Bonomi *et al.*, "Fog Computing and Its Role in the Internet of Things," *Proc. 1st ACM MCC Wksp. Mobile Cloud Computing*, Aug. 2012, pp. 13–16.

[2] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet of Things J.*, vol. 3, no. 6, Dec. 2016, pp. 854–64.

[3] S. Yi, C. Li, and Q. Li, "A Survey of Fog Computing: Concepts, Applications and Issues," *Proc. 2015 ACM Wksp. Mobile Big Data*, June 2015, pp. 37–42.

[4] S. Li, M. A. Maddah-Ali, and A. S. Avestimehr, "Coded MapReduce," *Proc. 2015 53rd Annual Allerton Conf. Commun. Control Computing*, Sept. 2015, pp. 964–71.

[5] ——, "Fundamental Trade-off between Computation and Communication in Distributed Computing," *Proc. 2016 IEEE Int'l. Symp.Info. Theory*, July 2016, pp. 1814–18.

[6] K. Lee *et al.*, "Speeding Up Distributed Machine Learning Using Codes," *Proc. 2016 IEEE Int'l. Symp. Info. Theory*, July 2016, pp. 1143–47.

[7] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Proc. 2004 6th USENIX Symp. Op. Sys. Design Implementation*, Dec. 2004, pp. 137–50.

[8] M. Zaharia *et al.*, "Spark: Cluster Computing with Working Sets," *Proc. 2010 2nd USENIX Workshop on Hot Topics in Cloud Comp.*, June 2010, pp. 10–10.

[9] M. A. Maddah-Ali and U. Niesen, "Fundamental Limits of Caching," *IEEE Trans. Info. Theory*, vol. 60, no. 5, May 2014, pp. 2856–67.

[10] ——, "Decentralized Coded Caching Attains Order-Optimal Memory-Rate Trade-Off," *IEEE/ACM Trans. Networking*, vol. 23, no. 4, Aug. 2015, pp. 1029–40.

[11] M. Ji, G. Caire, and A. F. Molisch, "Fundamental Limits of Caching in Wireless D2D Networks," *IEEE Trans. Info. Theory*, vol. 62, no. 2, Feb. 2016, pp. 849–69.

[12] "CodedTeraSort Implementations"; http://www-bcf.usc.edu/avestime/CodedTerasort.html, 2016, accessed Jan. 11, 2017.

[13] O. O'Malley, "TeraByte Sort on Apache Hadoop"; http://sortbenchmark.org/YahooHadoop.pdf, 2008, accessed Jan. 11, 2017.

[14] S. Li *et al.*, "Edge-Facilitated Wireless Distributed Computing," *Proc. 2016 IEEE GLOBECOM*, Dec. 2016, pp. 1–7.

[15] S. Li, M. A. Maddah-Ali, and A. S. Avestimehr, "A Unified Coding Framework for Distributed Computing with Straggling Servers," *Proc. 2016 IEEE Wksp. Network Coding Applications*, Dec. 2016.

Biographies

Songze Li [S'09] (songzeli@usc.edu) received his B.S. in electrical engineering from Polytechnic Institute of New York University (now NYU Tandon School of Engineering) in 2011, and his M.S. in electrical engineering from the University of Southern California in 2016. He is currently pursuing his Ph.D. as a research assistant in the Electrical Engineering Department, University of Southern California. His research interest is network information theory and its applications including interference management in wireless networks, and improving parallel/distributed computing using codes.

Mohammad Ali Maddah-Ali [S'03, M'08] (mohammad.maddahali@nokia-bell-labs.com) received his B.Sc. degree from Isfahan University of Technology, his M.A.Sc. degree from the University of Tehran, and his Ph.D. degree from the University of Waterloo. From 2008 to 2010, he was a post-doctoral fellow at the University of California, Berkeley. Since September 2010, he is with Bell Labs, New Jersey. He was the recipient of the 2015 IEEE Communications and Information Theory Societies Joint Paper Award and the 2016 IEEE Information Theory Society Paper Award.

A. Salman Avestimehr [M] (avestimehr@ee.usc.edu) is an associate professor in the Electrical Engineering Department of the University of Southern California. He received his Ph.D. (2008) and M.S. (2005) in electrical engineering and computer science from the University of California, Berkeley. His research interests include information theory, communications, distributed computing, and data analytics. He has received several awards, including the Communications Society and Information Theory Society Joint Paper Award, the Presidential Early Career Award for Scientists and Engineers (PECASE), and the National Science Foundation CAREER award.

RAINA: Reliability and Adaptability in Android for Fog Computing

Karthik Dantu, Steven Y. Ko, and Lukasz Ziarek

ABSTRACT

The ubiquity and universality of smartphones make them ideal fog devices to bridge edge devices and the cloud. However, to support a wide range of applications, as well as adhere to the resource constraints presented, the software stack on smart phones needs to be *reliable* and *adaptable*. We propose RAINA, an architecture to enable reliability and adaptability in Android. While our work is on Android, our ideas can easily be adapted to other mobile OSs. This article describes our software architecture, systems challenges, application challenges, and methods to address these challenges. We also discuss future work to allow smartphones to truly be at the center of the fog.

INTRODUCTION

The pervasiveness of cloud services, smartphones, wearables, Internet of Things (IoT) devices, and other embedded devices has led to the flow of computing, communication, and control toward the edge of the Internet. Small devices at the edge of the Internet (the "fog") are networked with each other and with cloud services, distributing computation and data management tasks across them to best serve end users. Envisioned fog systems span many domains such as personal entertainment, medical services, home automation, assistive robotics, and automotive navigation. Development of edge devices including household appliances (e.g., smart thermostats and robot vacuum cleaners), critical patient assistive devices (e.g., insulin pumps, cochlear implants, and cardio monitors), personal convenience gadgets (e.g., Microsoft Hololens and Samsung Gear VR), and personal health devices (e.g., heart rate monitors and smart watches) have all required the need for computing/communication/control that is closer to the user while still being connected to the Internet.

In all these applications, a user's smartphone, due to computational power, connectivity, and accessibility, is an ideal platform to use as a central hub that connects other edge devices and enable a rich set of features previously not possible. A modern smartphone is equipped with adequate computing, memory, and storage, making it possible to share some of its resources to other, more resource-constrained edge devices. For example, a smartphone may be leveraged to provide timely audio processing capability for a cochlear implant. In addition, a smartphone is equipped with many network interfaces, such as WiFi, Bluetooth, fourth generation (4G), near firld communication (NFC), and others, allowing it to connect to edge devices. For example, a smartphone can interact with a smartwatch via Bluetooth and an item tag via NFC. Lastly, a smartphone is almost always with the user, providing easy accessibility for edge devices in the vicinity of the user. For example, a smartphone can connect to all the smart bulbs in a room where the user is and display their control for the user to adjust. We believe that availability along with adequate resources as well as the connectivity *allows the smartphone to be the center of fog computing*.

It is due to these abilities that smartphones can enable a diverse set of user-centric applications. Several envisioned applications are potentially critical (monitoring a pace maker or cochlear implant), while others are not critical (browsing, taking pictures, social media), and still others are somewhere between (controlling the light bulb, Bluetooth headphones, augmented reality). To provide a combination of critical and non-critical services as required, the key properties required from the smartphone are *reliability* and *adaptability*. A system is said to be reliable if an application runs in a timely manner, uses resources as expected, and performs the desired set of actions. Providing such reliability requires a holistic approach that considers all layers of a system, including the underlying platform (the operating system [OS] and system libraries) as well as applications. Considering the underlying platform is necessary since reliability is primarily a platform property. If the OS and its system libraries do not have proper mechanisms to provide reliability, the whole system will not be reliable. In addition, considering applications is necessary since application programming with reliability in mind is not straightforward. Even if the underlying platform provides application programming interfaces (APIs) and system services that provide reliability guarantees, if application programmers do not leverage those APIs and services, users will not benefit from the platform-level reliability.

Shown in Fig. 1 is a visualization of several contemporary mobile/embedded software systems along the dimensions of reliability, adaptability, and deployability. Traditional real-time OSs such as VxWorks are very popular (high deployability) and reliable. However, runtime customization is very challenging in such an environment. This is

The ubiquity and universality of smartphones make them ideal fog devices to bridge edge devices and the cloud. However, to support a wide range of applications, as well as adhere to the resource constraints presented, the software stack on smartphones needs to be reliable and adaptable. The authors propose RAINA, an architecture to enable reliability and adaptability in Android.

The authors are with the University of Buffalo, State University of New York.

Digital Object Identifier:
10.1109/MCOM.2017.1600901

Hardware can vary, and applications may be required to be deployed on many different hardware platforms or migrate to newer platforms as devices become obsolete. To accommodate changing and evolving hardware platforms for fog applications, we need a framework that is adaptable.

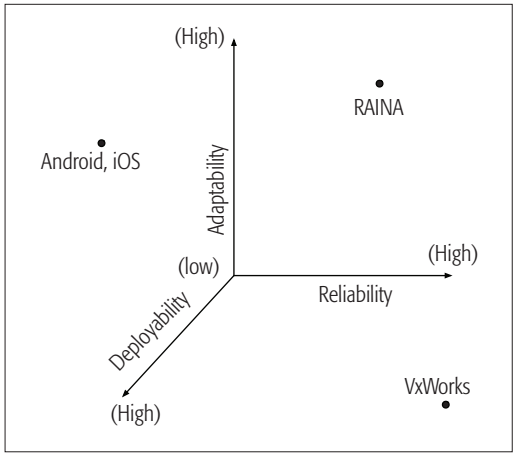


Figure 1. Comparison of embedded and sensing software systems with respect to RAINA.

very inconvenient to host on a smartphone without the ability to easily download apps from a store and use various services with the touch of a button. To satisfy this need, modern mobile OSs provide high adaptability and are definitely highly deployable. But they provide best effort service, which implies that they are relatively unreliable. This precludes them from use for critical applications such as healthcare or emergency messaging.

At the Reliable Mobile Systems (RMS) lab at the University of Buffalo, we are undertaking a multi-year effort to develop RAINA,[1] a reliable and adaptable software stack on smartphones. RAINA is being designed to support a variety of application domains, ranging from critical applications to non-critical applications. Figure 2 illustrates some example domains. RAINA is divided broadly into four projects: *RTDroid* [1, 2], *RA* [3], *BlueSeal* [4], and *Reptor* [5], which fall under two broad categories.

Systems Challenges: Mobile OSs have grown to be large pieces of software able to support a multitude of applications. However, most of these features are provided on a best effort basis. Several low-level constructs such as memory allocation and communication constructs do not provide priorities and timeliness guarantees, and therefore are unreliable. *RTDroid* is our effort to correct these problems and make Android, a popular mobile OS, reliable.

It is also hard to track resource usage by individual apps, making it challenging to enable efficient and fair use of different resources across applications. A battery is a scarce resource, and providing accurate usage information allows the system to make the most efficient use of the system. *Resource Accounting* (RA) [3] is our project to provide accurate resource accounting of asynchronous resources in Android. It provides OS and user-level constructs to accurately track I/O calls and attribute resource usage to individual apps. This allows for user policies for resource and energy usage to be implemented, fairly allowing for improved use of the smartphone.

Application Challenges: From the decades of past experience with real-time systems, it is now well understood that writing reliable applications is notoriously difficult. Although many APIs and system services do provide reliability guarantees for applications, leveraging them correctly without errors is a difficult task. To alleviate this problem, we are exploring a programming model similar to Android and integrating it with *RTDroid*, providing programmers with building blocks that are easy to leverage to construct reliable applications. A second challenge with real-time applications is their adaptability. We are investigating a specification mechanism coupled with virtual machine implementations that hide the complexities of achieving reliability from the programmer. In doing so, we leverage *BlueSeal*, our Android app analysis engine that is able to analyze Android constructs and their interactions, as well as *Reptor*, our Android app rewriting framework able to instrument existing Android apps. These tools allow us to simplify app instrumentation, thereby making apps adaptable to the user's needs.

Figure 3 illustrates our envisioned system stack for RAINA and its integration as a fog computing framework. *Reptor* and *BlueSeal* provide a powerful compile time framework for automatically adapting applications for RAINA, while RTDroid and RA provide necessary runtime support to provide guarantees to applications at runtime. This allows for reliable and predictable interactions with sensor and micro embedded devices, allowing RAINA to be used in the real-time domain. Robust Android-based connectivity allows for seamless integration to the cloud computing systems, making RAINA an ideal substrate to bridge low-evel sensor and embedded devices with high-level computing platforms like the cloud.

RAINA: Systems Challenges

Any platform that provides reliability requires two things: (a) a widely adopted infrastructure that can be deployed on a multitude of devices and (b) mechanisms that allow us to build required software components with necessary guarantees. Interactions within applications and between applications must be well defined to achieve correctness. For fog computing, this is even more crucial due to potentially evolving software systems. For example, mobile devices are highly adaptable, allowing their users to install, remove, update, and in some cases optimize software components. Mobile software components usually allow applications to dynamically leverage each other's exposed capabilities. To be able to reason about the correctness of deployed software, we need a communication mechanism that is *secure*. Since many applications must be able to meet their timing requirements, we need mechanisms that are *reliable*. Hardware can vary, and applications may be required to be deployed on many different hardware platforms or migrate to newer platforms as devices become obsolete. To accommodate changing and evolving hardware platforms for fog applications, we need a framework that is *adaptable*.

Timeliness of Operations

Current mobile OSs (Android, iOS, Windows Mobile, etc.) have little support for timeliness guarantees since they were designed for mobile devices and optimized for device mobility, user experience, and energy efficiency. There are two sources of unreliability on mobile platforms. First, the software stack as a whole does not have the

[1] RAINA is alternatively a short form for RAina Is Not Android.

concept of priorities. Thus, an application that requires a timely result may have to wait for an non-critical background task. This often shows up in the event-driven task communication infrastructure. The second source of unreliability comes from the bowels of the platform where services such as the garbage collector can preempt application code and lead to significant pauses. Using Android as our basis, we are incorporating mechanisms to the framework to provide timeliness to user apps under the RTDroid project (http://rtdroid.cse.buffalo.edu/).

The goal of RTDroid is to look for non-invasive ways to make Android suitable for computational tasks that have timeliness requirements. We focus on the programming model, but also address the underlying infrastructure. By non-invasive, we mean that existing Android applications should continue to work without changes on the platform, while new applications can be written in a style that is not too alien for Android developers. Our work builds on previous research that enforced strict isolation between computations [6], and studies on how to reduce memory management latencies [7], add priorities to core communication primitives [8], and add priorities in the lower levels of the Android stack [9]. The contribution of our work is a programming model for writing applications on Android-equipped devices [1] that can deliver soft real-time guarantees to applications that use it, while still allowing legacy code to run as before. More precisely, we aim to incorporate the following changes to the platform.

Declarative Timeliness: A declarative mechanism for programmers to specify the timeliness and resource requirements for their applications, without entangling such specifications in the application itself

Priority-Aware Communication: Specialized communication primitives that preserve the Android communication model, but provide programmers control over how components of potentially differing priority levels and with different timeliness guarantees communicate

Pauseless Memory Management: An implementation of our proposed constructs that internally leverages region-based memory management to avoid interference from the garbage collector for computations that require timeliness guarantees

Extended APIs: Extensions of existing constructs to specify required real-time behaviors and interactions

Efficient Resource Accounting and Usage

Given the multitude of applications envisioned to use a smartphone platform, it becomes paramount to efficiently use resources. To this end, we require accurate resource usage, accounting in the system for fair allocation as well as enforcement. Traditionally, CPU and memory formed these resources, and OSs were designed to efficiently use these resources. However, smartphones and other mobile devices are dominated by a multitude of network interfaces, input modalities, displays, sensors, and actuators. Most such devices are I/O devices, with large usage latencies making traditional OS accounting mechanisms inaccurate. Primarily, accounting for asynchronous resources — resources that are potentially used even when the requesting process is not running on the processor —is particularly challenging [10, 11].

Figure 2. Example application domains supported by RAINA.

To address this challenge, we propose RA [3], a general framework for accurate resource accounting in the OS as well as enhancements to user space daemons to provide access to this accounting information. It is based on our observation that *in order to reason about asynchrony, we need to relate a user-level request for a resource to the corresponding kernel-level request issued to the resource hardware*. Once we track these causal relations between user-level and kernel-level requests, we can accurately attribute a request and its actual resource use to its originating process.

We achieve this causality tracking with a general architecture composed of three core components: the *top-half*, which sits in the user-space-kernel boundary and monitors the interaction between user processes and the kernel; the *bottom-half*, which sits in the kernel-hardware boundary that monitors the interaction between the kernel and the hardware resources; and the *RA module*, which is a bookkeeping kernel module that associates interactions crossing the boundaries at the top-half and the bottom-half. Given this architecture, a device driver writer just needs to add simple timing measurement code in the device driver to enable this accounting.

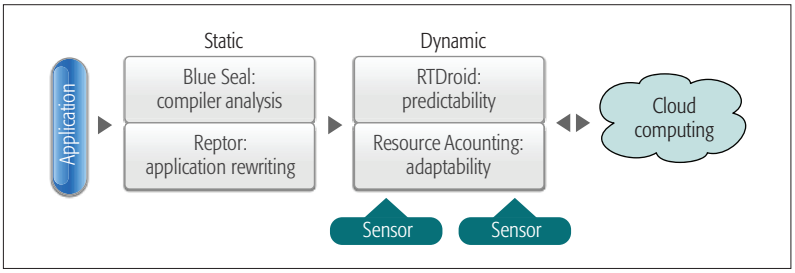


Figure 3. RAINA system stack.

Our prototype implemented in Linux demonstrates the feasibility and the benefits of our approach. We implement our proposed architecture in two subsystems, the sensor subsystem and the network subsystem. For the sensor subsystem, the top-half monitors system calls (e.g., `open`, `close`, `read`, `write`, and `ioctl`) as well as serial device access through `tty`. For the network subsystem, the top-half monitors system calls (e.g., `send` and `receive`) and sockets. With these cases, we show that our approach provides a more accurate method of resource accounting for asynchronous resources.

Our contributions are as follows:

- We identify asynchrony as the main challenge for accurate resource accounting for modern computing systems such as mobile devices, wireless sensors, and robotic systems.
- We propose a general architecture that accurately tracks asynchronous uses of various devices.
- We implement a prototype of the architecture with two subsystems (sensors and network interfaces) and demonstrate the benefits of our approach.

RAINA: Application Challenges

Although the systems we have described so far provide predictable primitives at the system level, they are not enough for developing predictable applications. This is because programmers need to utilize the primitives correctly in order to develop reliable software, which is known to be a difficult task. Thus, the second challenge we tackle is to develop tools to help programmers enable predictability in their applications. Our goal is to allow programmers who might not be familiar with the new primitives that our systems provide for predictability to write an app using existing mechanisms and APIs of Android. We then ask the programmer to provide a specification for the requirements and structure of the predictable (i.e., real-time) components. Using this specification, we analyze the original app's structure to see if it is *compatible* with respect to the specification, and synthesize a predictable variant of the app from the original app and specification. This allows automatic transformation of a regular Android app into a version of the app that uses our primitives for predictability.

Predictability Specification

Android applications are a collection of loosely coupled objects and services. Android provides a manifest configuration file to allow an application to declare constructs, runtime access permission, and which type of events one construct can receive. For example, it lists resources the application can access and defines mappings between `Intents` and their target constructs. This allows the programmer to statically express and limit the interactions of components. However, Android's specification does not provide a way to declare, either statically or dynamically, the runtime requirements of these objects and services. For predictable application development, it is desirable (if not necessary) to provide static bounds on the resources required by each component of the system, such that guarantees can be made about runtime performance and predictability. Once these bounds have been established, it is likewise desirable to enforce them at runtime, or at least detect when they have been violated, so that mitigation techniques can be put in place. For example, the manifest cannot express configuration of parameters that affect the timeliness of an application. Likewise, there is no mechanism to limit the *rate* of interaction through message passing or callbacks; nor is there the capability to express component priorities and memory allowances.

Since Android already has a default mechanism of statically declaring application configuration parameters, it is a natural mechanism for us to leverage for real-time configuration and specification as well. Our system reads all real-time parameters from the file, preallocates memory regions, and executes different tasks according to the specification. Expressing the system requirements and constraints via a declarative specification allows for the decoupling of configuration from the implementation. A direct consequence of this approach is that preallocation of components can be separated out from the application itself. Additionally, a VM can reject applications for which it cannot satisfy enumerated requirements. For example, verification guarantees the correctness of the application in two aspects.

Memory Boundary Checking: The total memory of a component should be equal to the sum of objects of its persistent memory, its release memory, and the release memories of all its sub-components.

Channel Overflow Checking: The incoming message rates should not exceed the message processing rates for each channel.

Analysis of Android Apps

To reach our goal of synthesizing predictable mobile apps, we must first be able to understand app logic and structure and the correspondence of this code to the provided specification. We note, however, that we do not verify that the logic and structure will meet the requirements of the specification; instead, we analyze if the app logic can be structured to mirror the specification (i.e., creating a real-time-aware service from a standard service or set of services). To accomplish this, we plan to leverage our BlueSeal analysis engine [4], which can analyze the components of an Android app as well as the interactions between components.

The BlueSeal analysis engine will take in an RTDroid manifest as a specification as well as a standard app. This manifest will include information about all real-time components in the system and their specific requirements. Additionally, this manifest will include the interactions the com-

ponents can perform and limitations on such interactions. BlueSeal will analyze the app, its components, and interactions between components to see if the app structure adheres to the specification. Lastly, if the app's structure satisfies the specification, we leverage the Reptor re-writing tool [5] to synthesize RTDroid components from their Android counterparts. This effectively re-writes an Android app to an RTDroid app.

Pre-Allocation of Resources

Since analyzing code to see if it adheres to real-time constraints provided in a specification is difficult and an open research problem, our proposed system instead leverages RTDroid's ability to limit resource usage through pre-allocation and predictable resource management at runtime. To take full advantage of the static configuration defined by the manifest, RTDroid employs a multi-stage configuration and boot process. Figure 4 shows the entire system boot process, starting with compile time configuration processing, which is divided into five logical stages. During compile time, the Fiji VM compiler is responsible for parsing the manifest XML file. The compiler emits configuration classes for each of the configured objects in the manifest. It also emits the `Configuration Object`, which provides a unique handle to all configuration classes and objects needed by the boot sequence. Booting is accomplished in four steps. First, the VM performs its own startup process and instantiates the `Configuration Object`. Once this is complete, the VM hands off control to the RTDroid system. RTDroid is responsible for initializing all system services. Once the system services have been initialized, RTDroid initializes all Android components required by the application running on RTDroid. Information for all components and system services is defined in the manifest. Lastly, an intent (conceptually an asynchronous event) is delivered to the application, which triggers the application execution.

Conclusion

Smartphones and portable computing devices bridge cloud computing technologies with embedded devices and IoT, thereby creating mobile fog computing. Future platforms and applications in the mobile fog demand varying degrees of predictability in the apps deployed on these devices. The RMS lab is working to develop necessary technologies for ensuring predictable fog computing.

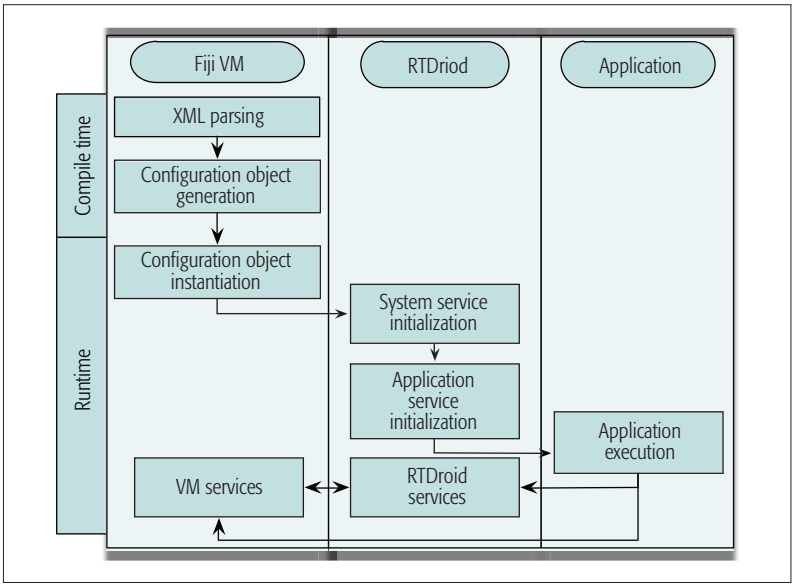


Figure 4. RTDroid configuration and boot sequence.

References

[1] Y. Yan *et al.*, "Making Android Run on Time," accepted for publication, *IEEE Real-Time and Embedded Technology and Applications Symp.*, 2017.

[2] Y. Yan *et al.*, "RTDroid: A Design for Real-Time Android," *IEEE Trans. Mobile Computing*, vol. 15, no. 10, Oct., 2016.

[3] F. Ghanei *et al.*, "OS-Based Resource Accounting for Asynchronous Resource Use in Mobile Systems," *Proc. 2016 ACM Int'l. Symp. Low Power Electronics Design*, pp. 296–301.

[4] F. Shen *et al.*, "Information Flows as a Permission Mechanism," *Proc. 29th IEEE/ACM Int'l. Conf. Automated Software Engineering*, 2014.

[5] T. Ki et al., "Reptor: Enabling API Virtualization on Android for Platform Openness," to appear, *Proc. 15th Int'l. Conf. Mobile Systems Applications Services*, 2017.

[6] C. Maia, L. Nogueira, and L. M. Pinho, "Evaluating Android OS for Embedded Real-Time Systems," *Proc. 6th Int'l. Wksp. Operating Systems Platforms Embedded Real-Time Applications*, Brussels, Belgium, 2010, pp. 63–70.

[7] I. Kalkov *et al.*, "A Real-Time Extension to the Android Platform," *Proc. 10th ACM Int'l. Workshop Java Technologies Real-Time Embedded Systems* 2012, pp. 105–14.

[8] I. Kalkov, A. Gurghian, and S. Kowalewski, "Predictable Broadcasting of Parallel Intents in Real-Time Android," *Proc. 12th ACM Int'l. Workshop Java Technologies Real-Time Embedded Systems*, 2014, pp. 57:57–66.

[9] Y. Yan *et al.*, "Real-Time Android with RTDroid," *Proc. 12th ACM Int'l. Conf. Mobile Systems Applications Services*, 2014.

[10] H. Zeng *et al.*, "ECOSystem: Managing Energy as a First Class Operating System Resource, *Proc. 10th ACM Int'l. Conf. Architectural Support for Programming Languages and Operating Systems*, 123–32.

[11] M. Dong, T. Lan, and L. Zhong, "Rethink Energy Accounting with Cooperative Game Theory, *Proc. 20th ACM Annual Int'l. Conf. Mobile Computing and Networking*, pp. 531–42.

Biographies

Karthik Dantu (kdantu@buffalo.edu) is an assistant pofessor in the Department of Computer Science and Engineering at the University of Buffalo, State University of New York. Previously, he received his Ph.D. from the University of Southern California and was a postdoctoral fellow at Harvard University. His research interests are in mobile systems and robotics.

Steven Y. Ko (stevko@buffalo.edu) is an assistant professor in the Department of Computer Science and Engineering at the University of Buffalo. His research interest spans systems and networking, with the current focus on mobile systems. He received his Ph.D. from the University of Illinois at Urbana-Champaign in computer science.

Lukasz Ziarek (lziarek@buffalo.edu) received his B.S. degree in computer science from the University of Chicago and his Ph.D. degree in computer science from Purdue University. He is an assitant professor in the Department of Computer Science and Engineering at the University of Buffalo. His research interests are in programming langauges and real-time systems.

5G Radio Access Network Design with the Fog Paradigm: Confluence of Communications and Computing

Yu-Jen Ku, Dian-Yu Lin, Chia-Fu Lee, Ping-Jung Hsieh, Hung-Yu Wei, Chun-Ting Chou and Ai-Chun Pang

The authors describe recent advances in fog radio access network (F-RAN) research, hybrid fog-cloud architecture, and system design issues. Furthermore, the GPP platform facilitates the confluence of computational and communications processing. Through observations from GPP platform testbed experiments and simulations, they discuss the opportunities of integrating the GPP platform with F-RAN architecture.

ABSTRACT

Cloud-based wireless networking system applies centralized resource pooling to improve operation efficiency. Fog-based wireless networking system reduces latency by placing processing units in the network edge. Confluence of fog and cloud design paradigms in 5G radio access network will better support diverse applications. In this article, we describe the recent advances in fog radio access network (F-RAN) research, hybrid fog-cloud architecture, and system design issues. Furthermore, the GPP platform facilitates the confluence of computational and communications processing. Through observations from GPP platform testbed experiments and simulations, we discuss the opportunities of integrating the GPP platform with F-RAN architecture.

INTRODUCTION

When it comes to the development of the telecommunication system nowadays, the concept of centralization is always mentioned. The concept of centralized architectures such as the baseband unit (BBU) pool of the cloud radio access network (C-RAN) and third party cloud computing center services have been widely discussed. The concept of moving the computing, storage, and networking functions from the local end to the cloud enables the operators to manage the system in an efficient and energy-saving way.

Meanwhile, the increasing number of smart Internet of Things (IoT) devices and emerging low-latency applications have attracted much attention in the industry [1, 2]. In an IoT network, loading between massive devices and cloud servers might lead to heavy loading on backhaul [3]. In addition, latency to a centralized site is an issue for delay-sensitive applications. One of the solutions is to distribute a part of the computing services from cloud computing centers to the network edge. Such deployment is also known as mobile edge computing (MEC) or fog computing. By implementing the radio access network (RAN) with the fog computing paradigm, some of the control and data plane functions can be processed not only in the BBU pool but also at a local BBU or even remote radio heads (RRHs). The idea was named fog RAN (F-RAN) and was first proposed at the Next Generation Mobile Networ (NGMN) Forum in June 2014 [2]. With multiple computational tiers in F-RANs, the applications could be handled at a third party cloud computing server, a network edge node, or even local BBUs.

Several performance metrics have been identified for fifth generation (5G) wireless systems. Among them, low end-to-end latency will be a challenge [4]. As an end-to-end system is composed of several protocol layers and system components, the delay value in each protocol layer and system component needs to be handled carefully to minimize the end-to-end latency. From the radio interface to baseband processing to higher-layer protocols to computing, everything needs to be designed to support low latency. To meet the requirements of mission-critical applications in 5G mobile networks, a key RAN design principle is to embrace the fog paradigm.

TRENDS IN RAN ARCHITECTURE

CONFLUENCE OF FOG AND CLOUD

There are two system design paradigms: cloud and fog. The cloud-based design paradigm applies centralized resource pooling to achieve efficient resource utilization. Leveraging the advantages of both cloud-based and fog-based designs, a hybrid architecture that integrates the C-RAN and the F-RAN has been proposed.

In [5], the fog-cloud integrated RAN, which combines the C-RAN/heterogeneous C-RAN (H-CRAN) and the F-RAN, was proposed. In this architecture, there are four types of clouds: global centralized communication and storage cloud, centralized control cloud, distributed logical communication cloud, and distributed logical storage cloud. Here, we give a hybrid architecture example as shown in Fig. 1. The global centralized communication and storage cloud in the centralized BBU pool provides flexible management of radio signaling processing and resources in subordinate radio access points such as RRHs and fog-computing-based access points (F-APs).

In F-RANs, distributed logical communication and storage clouds are used. These clouds are composed of edge devices, such as RRHs, F-APs, and user equipments (UEs), which support direct device-to-device (D2D) communication with other UEs. Application processing can be performed in the edge devices to reduce latency and

Digital Object Identifier:
10.1109/MCOM.2017.1600893

The authors are with National Taiwan University; Hung-Yu Wei is the corresponding author.

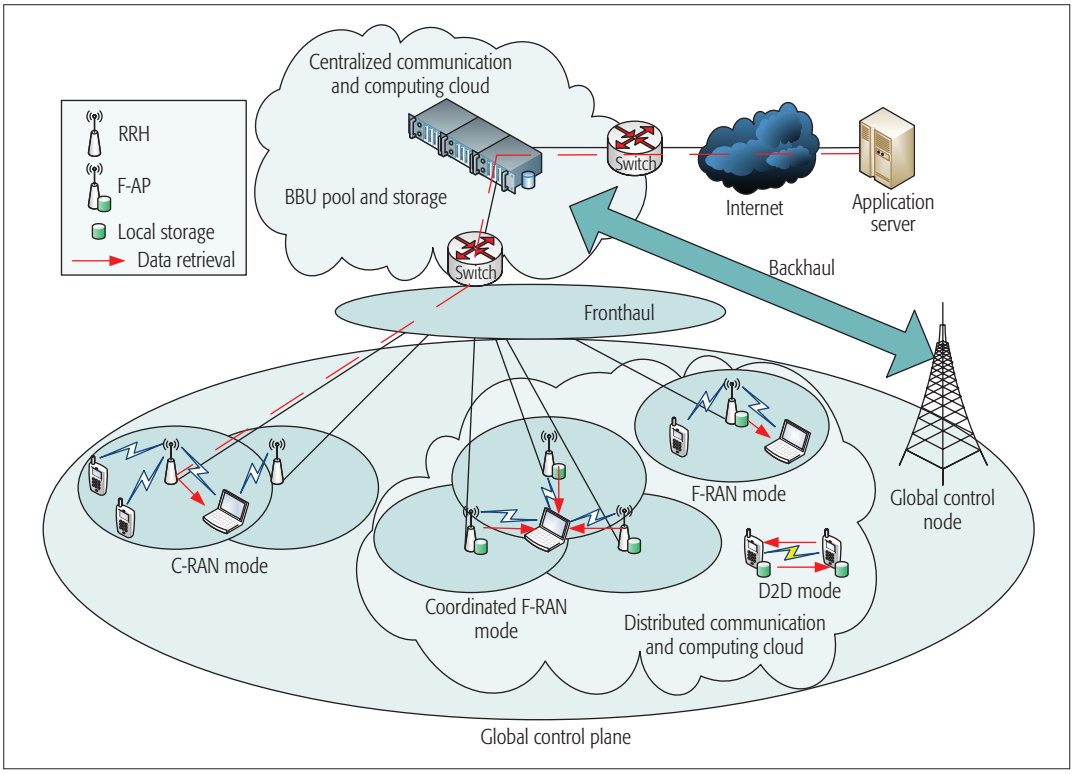


Figure 1. Fog-cloud integrated RAN architecture.

The F-RAN can be considered as a wireless networking platform that closely combines communications and computing operations. In each of the system design dimension (communications and computing), the architecture could be either fog-based, cloud-based, or hybrid. To integrate the computing functionality to 5G communications networks, two approaches could be adopted: loosely-coupled and tightly-coupled.

traffic loading to the BBU pool. Furthermore, adjacent edge devices can be interconnected to provide coordinated and cooperative functionalities among the devices. Thus, each edge device may exchange application data, and then relay the caching data to UEs. Each network topology (e.g., mesh, tree, or star topology) has its own pros and cons. In [2], a logical star topology, where a master F-AP distributes computing-intensive application tasks to its slave F-APs and summarizes the processing results, is used. The performance-cost trade-off is affected by the number of participating F-APs, the amount of assigned tasks for each participating F-AP, and the status of F-APs.

The switch/selection of the RAN and application processing among different clouds depends on the statuses of the edge devices and the characteristics of application tasks. On the basis of the four clouds in the fog-cloud integrated RAN, four transmission modes to be selected for a UE are D2D, local coordination, global centralized, and centralized control modes, respectively. With a similar architecture in [6], the FogNet-HCRAN architecture proposed is proposed based on the harmonization of cloud-based H-CRAN and fog-based FogNet. Design issues such as mobility management, caching in the application layer, and resource access control are studied. To allocate the proper processing resources or switch a UE between the H-CRAN and FogNet, system optimization is conducted. In [7], recent work on the performance analysis of caching and radio resource allocation in F-RANs is discussed.

Confluence of Computing and Communications

The F-RAN can be considered as a wireless networking platform that closely combines communications and computing operations. In each of the system design dimensions (communications and computing), the architecture could be either fog-based, cloud-based, or hybrid. To integrate the computing functionality to 5G communications networks, two approaches could be adopted: loosely coupled and tightly coupled.

Mobile Edge Computing: The European Telecommunications Standards Institute's (ETSI's) MEC effort [8] could be considered as a loosely coupled approach to integrate computing with communications. On the MEC platform, the new computing resource is utilized as an infrastructure for executing mobile edge applications. The mobile edge host on the platform manages the computing resource by supervising the status of the mobile edge applications, for example, controlling the registration and life cycle of each mobile edge application. To efficiently make use of the computing resource, the MEC platform hosts both the network function and mobile edge applications on the virtualized infrastructure that provides computing, storage, and network services [3, 9]. Although the MEC resource and RAN resource are managed separately, as a benefit of loosely coupled architecture, an MEC host is able to request some RAN resources such as bandwidth for mobile edge applications with higher prioritization requirements. RAN communications and computing processing at network edges could be integrated with MEC to provide desirable features for applications. For example, an application that has a low latency requirement can choose to implement the rendering pipeline either in a mobile edge application running on the mobile edge host or directly on the client device [10]. For instance, mobile edge applications can use the information provided by the radio network information to improve the quality

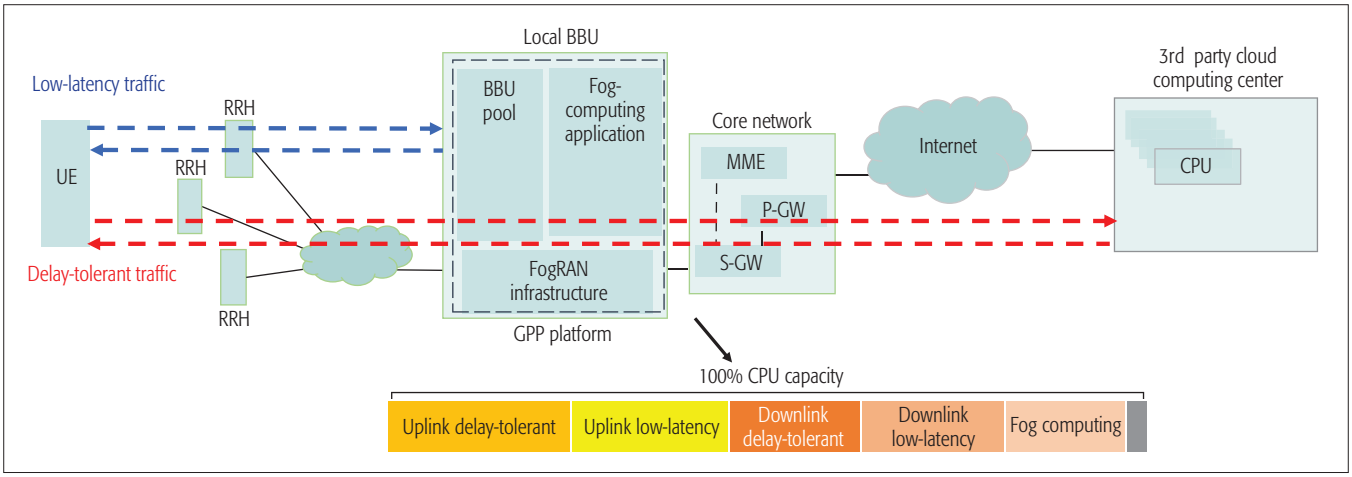


Figure 2. GPP-platform-based F-RAN.

of service. Another example is where frequently accessed objects are stored in the MEC in order to reduce the downloading time of popular contents in the local area [11].

GPP Platform: To integrate computing with communications, we could apply the tightly coupled approach for the next generation RAN design. The proliferation of the general-purpose processor (GPP) platform in wireless networking systems will lead to another paradigm shift. Typically, conventional RAN equipment is built with specific and dedicated processing resources. Instead, in CPU-based platforms or integrated CPU/field programmable gate array (FPGA) platforms, general and shared processing resources are available. In the emerging CPU or CPU/FPGA-based RAN platform, processing resources could be utilized for both communications and computing. As a result, efficient utilization of processing resources to meet the communications and computing requirements becomes an issue worthy of discussion.

One approach is to design F-RAN architecture based on the GPP platform [12]. The proposed scheme focuses on how to configure the computational power of the BBU pool to serve both the applications and communications processing. In the conventional setting of a BBU pool, there are separate processing units responsible for application and RAN processing. In contrast, in the GPP platform F-RAN, running all processes on the same computing unit enables the system to manage the computing resource in a more efficient manner. The shared resource in the GPP platform could be used for communications and computing. In an example of serving mixed delay-sensitive (DS) and delay-tolerant (DT) flows, as shown in Fig. 2, the CPU resource is divided for uplink DS transmission, uplink DT transmission, downlink DS transmission, downlink DT transmission, and application processing for MEC.

The GPP-based F-RAN includes three tiers: the global centralized cloud, the distributed local cloud located in the subordinate radio access points, and the distributed local cloud constructed among the UEs. Each tier can apply hybrid computational architectures for application and RAN processing with the GPP platform. On the other hand, the processing of the control plane can be centralized or localized in two tiers, which are the global centralized cloud and the distributed local cloud located in the subordinate radio access points. The distributed local clouds provide for low-latency signaling; for example, rapid intra-site handover for low latency traffic. The global centralized cloud, when configured for hybrid processing, combines the control plane/user plane computations for both cost savings (e.g., sharing the same BBU) and resource savings (e.g., saving equipment energy during off-peak times). The global centralized cloud may also be configured for dedicated processing for handling control plane messages.

Design Issues in F-RANs

Traffic Forwarding for an Integrated Fog-Cloud System

To support diverse applications in 5G systems, the concept of network slicing is proposed. Multiple logical network slices for different applications can be supported by a single but integrated network. For example, the network can collect data from IoT devices and simultaneously provide streaming service for mobile users. There would be various types of applications running over the same network, creating different kinds of traffic flows. Each traffic flow has its own requirement on latency. Some applications, such as vehicle and gaming applications, possess the low-latency requirement. On the other hand, relatively high latency would cause less effect on services like web browsing activities. Under the multiple computational tiers in a hybrid fog-cloud architecture, applications with distinct latency quality can be served by appropriate computing tiers. In this mixed traffic scenario, a differentiation mechanism needs to be applied to differentiate low-latency flows and delay-tolerant flows.

As shown in Fig. 2, by allocating the computational resource from the cloud computing center to delay-tolerant traffic and assigning the local BBU computing tier to low-latency requirement flows, the proposed F-RAN architecture is able to handle mixed flows of delay-tolerant and delay-sensitive application flows. The computing resource in the local BBU is limited. It serves not only as a BBU for wireless communications, but also as a fog computing server. Both delay-tolerant and delay-sensitive traffic flows consume the baseband computing power of the local BBU. As

a result, some of the low-latency services might not be served at the local BBU due to the lack of computing power, and need to be forwarded to back-end computing tiers.

To further analyze the cost of computing resource in the local BBU for either delay-tolerant or delay-sensitive traffic flows, capacity regions for each application should be considered. In the local BBU, all types of flows consume the same baseband processing resources. However, once the delay-sensitive application is allocated local fog computing service, the application would consume extra computational resources in the local BBU and therefore influence the capacity region of the application.

Note that for the local BBU, the cost of serving fog computing depends on the types of the applications. For example, for applications such as AR and video delivery with caching, AR consumes more fog computing resource when executing the application locally, while video delivery costs more communication resource due to downlink video transmission. Therefore, the capacity regions for both applications will not be the same. Based on the application profile, which includes the required downlink communication resource, uplink communications resource, and computational resource, we could derive the capacity region for the application. If the capacity regions for each application are derived, the RAN will have more information for admission control for a new application flow and dynamic traffic forwarding decisions.

Caching

In addition to the communication and computing design considerations, caching is the third dimension of design consideration in F-RANs. Caching has been identified as an important aspect by bringing storage functionality in network edges. Caching can improve quality of experience (QoE) of the consumers in F-RANs, especially in delay-sensitive content retrieval. For example, the users of online social networks tend to value contents highly recommended by friends, which can be cached beforehand [13]. On the other side, the download time of videos can be reduced by local content caching at the mobile edge host. Nowadays, high-resolution video can be played on handheld devices. After the request is identified by the content caching application, the user traffic will go through the video compression and video analytics application before it is delivered to the end user [8].

Interworking between F-RAN and User Devices

In the fog networking system, computing tasks could be handled by servers, network nodes, and user devices. Dividing a computing task between cloud servers, network edge nodes, and UEs needs to consider the bandwidth consumption, computation load requirement, and delay. For example, wearable devices with low levels of processing and storage capacities could migrate some application processing to nearby edge nodes, which could be co-located with a base station or WiFi AP. Meanwhile, some powerful mobile nodes such as laptops and high-end smartphones could handle more computing tasks in the device and migrate applications to edge nodes occasionally. On the other hand, some proximal user devices might collaboratively handle some application processing tasks. In such a case, D2D communications might facilitate the proximal collaboration between user devices. A hybrid model that integrates computing task migration between a user device, a fog node at the network edge, and a cloud server could provide flexible system operation. In a given F-RAN network-centric topology, the network may consist of fog nodes on the network edge, but the user devices do not provide for fog computing. Network-edge fog nodes might also be combined with cloud computing nodes, whereas for a user-provided F-RAN topology, F-RAN architecture might include active engagement of users. User devices might be active in fog computing tasks. In addition, they may collaborate with other user devices.

Security

Although the security issue is critical for the development of F-RAN systems, there is little discussion on the topic. By moving the RAN resources from the clouds to the edge, F-RANs will encounter some security concerns that might not occur in the conventional scenarios with centralized clouds only. Compared to the global centralized cloud RAN architecture, the F-RAN is more vulnerable to threats from malicious attacks on the system [1]. Under the F-RAN deployment, the distributed fog computing units might not be able to detect an attack due to the lack of global information of the whole network. In an even worse scenario, due to the fact the authentication policy is enforced on the edge end of the network, instead of in the cloud, the attacker could easily get permission to access the network through the end gateways and remain undetected by the global cloud or firewall.

The article [14] further analyzes the security issue. Moreover, it shows an example on how a man-in-the-middle attack can affect F-RANs. A stealth test has been done in a realistic environment, and the result shows that under the distributed F-RAN architecture, this kind of attack is not easy to detect.

On the other hand, the F-RAN still has some advantages in enhancing network security. For end devices that are equipped with constrained computing resource, the edge end computational tiers in F-RANs can serve as a security and virus scanning IoT supervisor and respond quickly if threats are detected [1].

Observations from an Experimental Testbed

To further examine the feasibility of the proposed F-RAN system, we have done a series of experiments on a realistic RAN system. We observed the CPU behavior of a local BBU in a testbed by measuring the CPU loading. The local BBU testbed is an 8-core GPP platform. Four cores are used for basic BBU functions and other operating system processes. During the observation, we first constructed a downlink or uplink channel between the core network and the UE end. Then we used iperf to generate traffic and monitor the CPU loading under different loading conditions.

We first measured the background CPU activity without running BBU-related processes on the platform. As the asterisk line of Fig. 3a shows, there are some minor CPU operations due to

Compared to the global centralized cloud RAN architecture, the F-RAN is more vulnerable to threaten from malicious attack to the system. Under the F-RAN deployment, the distributed fog computing units might not be able to detect an attack due to the lack of global information of the whole network.

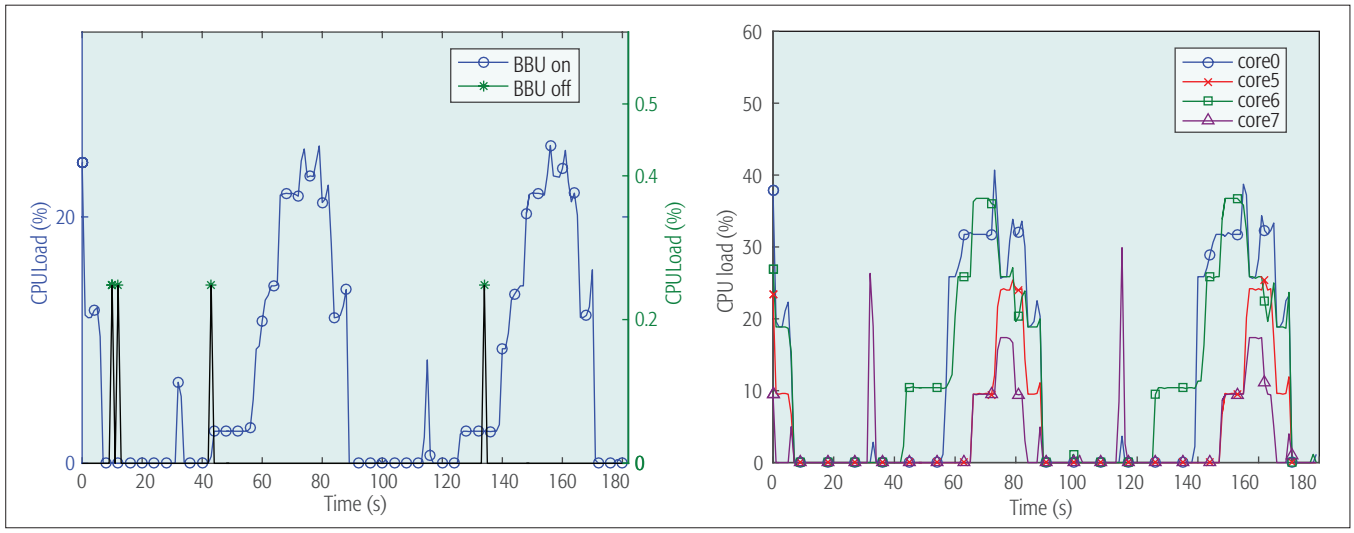


Figure 3. a) CPU load with 0M traffic; b) the load of each core with 0M traffic.

the background process. The circle line of Fig. 3a shows the measurement result after turning on the local BBU: periodic CPU activity was observed. The behavior of each CPU core is periodically recorded in Fig. 3b. When we started to transmit data, it could be observed that besides the periodic background CPU activity measured in the previous experiment, there is additional activity in CPU loading. The period of the background CPU load remained the same regardless of the value of the traffic loads. However, the extra CPU load rose with the increase of the traffic loads. As an example, Fig. 4a depicts the measurement of background CPU activities plus the extra CPU load based on 10 Mb/s downlink transmission.

To investigate the relationship between traffic load and CPU load, we changed the downlink and uplink transmission loads and recorded the increase of the computing resource usage in the local BBU. For downlink transmission, the result is shown in Fig. 4b. The circle line depicts the actual throughput of the connection, and the asterisk line shows the CPU load usage caused by different downlink traffic loading. Figure 4c shows another result under uplink transmission. A linear growth of the CPU load consumption could be observed for both downlink and uplink transmissions. Each type of transmission had its own CPU-traffic load ratio.

Based on the result in Fig. 4, we could predict the local BBU CPU load based on the traffic load information. Combined with the observed result in the first experiment, it would be possible to predict the remaining BBU computational resource for fog computing service. As traffic demand from diverse applications varies in the time and spatial domains, adaptive system operation is needed in a heterogeneous wireless network [15]. With a better understanding of the traffic-CPU loading relationship, context-aware policies could be applied for GPP-based F-RAN operations. The system could allocate computing resource to delay-sensitive and delay-tolerant applications wisely.

Observations from Traffic Forwarding Policies in an Integrated Fog-Cloud System

The proposed F-RAN architecture can be modeled as a queuing system, as illustrated in Fig. 5. Application traffic generated by UEs is categorized as delay-sensitive (DS) and delay-tolerant (DT). DS traffic has stringent latency requirements and needs prompt computation at the local application server. On the other hand, DT traffic could tolerate longer delay and is served by the cloud server. It is assumed that the total computational resource is fixed at an F-RAN node and could be dynamically allocated to baseband processing and application-dependent computation. Two F-RAN policy controllers are required for efficient resource utilization, as also shown in Fig. 5. The routing controller determines whether a DS request should be routed to the local or cloud

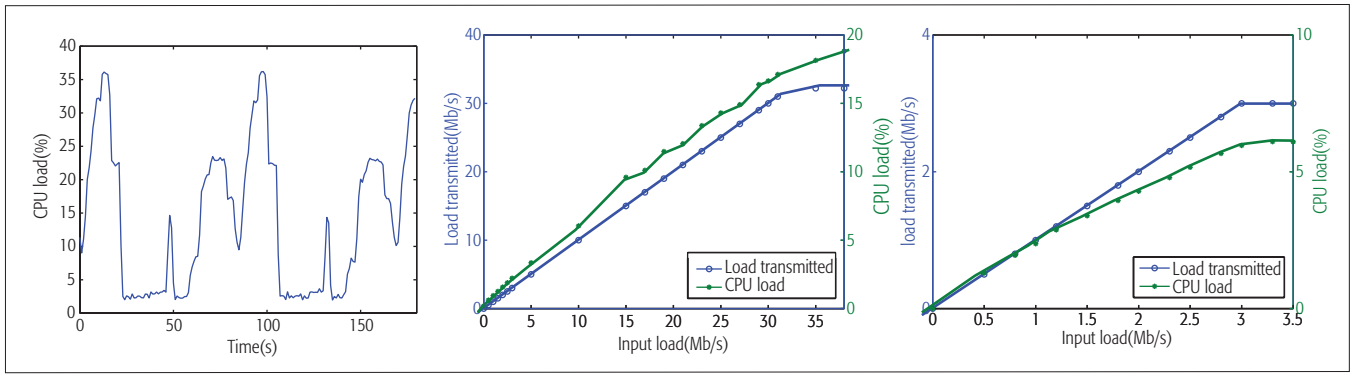


Figure 4. a) CPU load with 10Mb/s downlink transmission; b) downlink load; c) uplink load.

server, based on the status of the computational resource and queue size of the APP queue. In principle, the amount of DS traffic routed to the APP should avoid overflowing at the APP queue and overwhelming the baseband server. The processing prioritization controller determines whether or not service differentiation between uplink and downlink DS requests should be applied to the baseband processing server. Here, the downlink DS requests are referred to as the DS traffic that has completed its services at the local application server. A possible option would be that downlink DS traffic is granted higher priority (e.g., to be the head of the DS queue in Fig. 5). By doing so, the latency of DS traffic could be further reduced. One could expect a trade-off between end-to-end latency of DS traffic and blocking probabilities of uplink DS and DT traffic regardless of resource allocation, routing, and processing prioritization policies.

A simulation is conducted to evaluate the performance of the proposed F-RAN, based on the aforementioned queuing system. A very simple routing policy is adopted where uplink DS traffic is routed to the local application server as long as the APP queue is not full. In order to avoid long queuing delay, the maximum APP queue size is set as 3 (requests). Otherwise, the queuing delay at the APP queue would easily diminish the benefit of local application processing. A first-in first-out (FIFO) policy is applied to the DS queue so that uplink and downlink DS requests are treated equally. An alternative policy is giving downlink DS requests higher priority by placing the requests at the head of the DS queue. Although the downlink DS requests would benefit from this policy, the uplink DS traffic would also face a longer queuing delay and a higher blocking probability. As a result, the end-to-end latency may not necessarily be reduced. Finally, it is assumed that there are a total of 14 computation units (e.g., 14 CPU cores) at the fog node where 7, 4, and 3 are reserved for DS baseband processing, DT baseband processing, and local application, respectively. The allocation of computation resource to different services at the fog node has a significant impact on the end-to-end latency performance. For example, if r percent of uplink DS traffic is routed to the local application server, the computation units in the DS baseband server should be at least (100+r) percent as large as in the local application server to avoid unstable operation.

Figure 6 shows the end-to-end latency and successful rate of DS traffic. The end-to-end latency of a DS request includes both uplink and downlink queuing and processing delay in the baseband server, and the queuing and processing delay at the local application or cloud server, depending on the route the request uses. The successful rate is defined as the percentage of the DS requests that are served successfully and received by UEs. The average latency in the figure is the weighted average of the latency experienced by DS requests through the local application server and the latency through the cloud server, which are also shown in the figure. The results show that although under the proposed F-RAN architecture the DS traffic being routed through the cloud server experiences higher end-to-end delay compared to the traditional C-RAN, that being routed through the local application server experiences much shorter delay. As a result, the average latency is reduced compared to the C-RAN architecture for most of the traffic loads. For the cases where the traffic load is high, computation resource allocated to the baseband server does not suffice to serve baseband processing, let alone the local application computation. Therefore, the average latency is higher than that under the C-RAN architecture. The observations verify that a dynamic routing policy that adapts to the resource utilization is key to the end-to-end latency reduction under the proposed F-RAN. Figure 6 also shows that some uplink DS traffic, even only 6 percent in the worst cases, is dropped under the proposed F-RAN, compared to 0 percent (i.e., 100 percent successful) under the C-RAN. In our simulation, 11 (7 + 4) computation units are allocated to the DS baseband server, and all uplink DS traffic is routed to the cloud server. As a result, no DS request is dropped. Under the proposed F-RAN, both uplink and downlink DS requests are

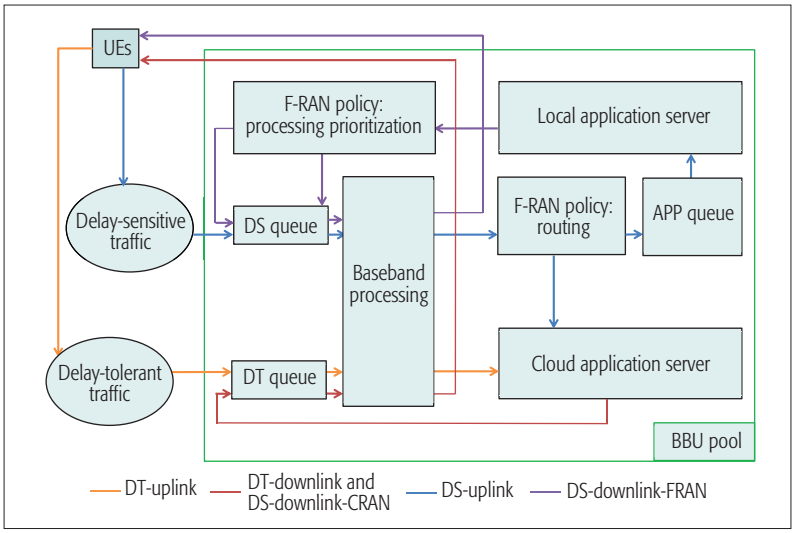


Figure 5. Model of the F-RAN architecture.

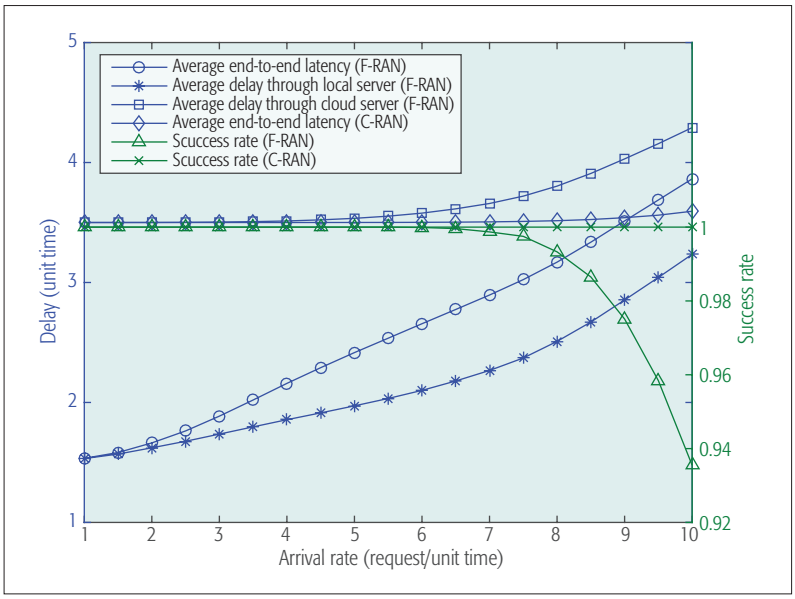


Figure 6. Delay and success rate of DS traffic.

Based on the experimental testbed observations, adaptive resource allocation and traffic forwarding in F-RANs has the potential to provide better QoE for diverse applications. Security issues and interworking design between the F-RAN and user devices are also important future research directions.

subject to being blocked at the baseband server as only seven units are allocated. The trade-off between the end-to-end latency and the success rate results from the allocation of computation resource to the baseband processing and application processing. One would need to make such a trade-off via user-defined or application-dependent utility function, which can easily be integrated into the proposed F-RAN architecture.

Conclusions

The F-RAN is promising for low-latency operation in 5G. Hybrid architecture can leverage the respective advantages of the F-RAN and the C-RAN. The integrated fog-cloud wireless design system provides efficient operation to support diverse applications. The emerging GPP platform enables the F-RAN for flexible operation in communications, computation, and storage. Based on the experimental testbed observations, adaptive resource allocation and traffic forwarding in F-RANs have the potential to provide better QoE for diverse applications. Security issues and interworking design between the F-RAN and user devices are also important future research directions.

Acknowledgment

The authors are grateful for the funding support from Foxconn and the Ministry of Science and Technology (MOST) of Taiwan under grants 103-2221-E-002-086-MY3, 105-2221-E-002-014-MY3, 105-2221-E-002-144-MY3.

References

[1] M. Chiang and T. Zhang, "Fog and IoT: An Overview of Research Opportunities," *IEEE Internet of Things J.*, vol. 3, Dec. 2016, pp. 854–64.
[2] Y. Shih *et al.*, "Enabling Low- Latency Applications in Fog Radio Access Networks," *IEEE Network*, vol. 31, no. 1, Jan. 2017, pp. 52–58.
[3] R. Munoz *et al.*, "The CTTC 5G End-to-End Experimental Platform : Integrating Heterogeneous Wireless/Optical Networks, Distributed Cloud, and IoT Devices," *IEEE Vehic. Tech. Mag.*, vol. 11, Jan. 2016, pp. 50–63.
[4] 3GPP, "Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology TR 22.862 V14.0.0," June 2016.
[5] M. Peng *et al.*, "Fog-Computing-Based Radio Access Networks: Issues and Challenges," *IEEE Network*, vol. 30, no. 4, July 2016, pp. 46–53.
[6] S. Hung *et al.*, "Architecture Harmonization between Cloud Radio Access Networks and Fog Networks," *IEEE Access*, vol. 3, Dec. 2015, pp. 3019–34.
[7] M. Peng and K. Zhang, "Recent Advances in Fog Radio Access Networks: Performance Analysis and Radio Resource Allocation," *IEEE Access*, vol. 4, Aug. 2016, pp. 5003–09.
[8] ETSI MEC ISG, "Mobile Edge Computing (MEC); Framework and Reference Architecture GS MEC 003 V1.1.1," Mar. 2016.
[9] ETSI MEC ISG, "Mobile Edge Computing (MEC); Technical Requirement GS MEC 002 V1.1.1," Mar. 2016.
[10] A. Neal and NGMN P1 WS1 E2E Architecture Team, "Edge Computing V0.4.2," July 2016.
[11] OpenFog Consortium Architecture Working Group, "OpenFog Architecture Overview OPFWP001.0216," Feb. 2016.
[12] Y. Ku, D. Lin, and H. Wei, "Fog RAN over General Purpose Processor Platform," *IEEE VTC-Fall 2016*, Sept. 2016.
[13] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 82–89.
[14] I. Stojmenovic and S. Wen, "The Fog Computing Paradigm: Scenarios and Security Issues," *Comp. Sci. and Info. Systems*, Sept. 2014.
[15] G. Lin and H. Wei, "Flexible 5G M2M Network Access with Cognitive RAN: Survey and Design Principles," *Proc. IEEE Conf. Standards Commun. Networking 2015*, Oct. 2015.

Biographies

Yu-Jen Ku (yuku@eng.ucsd.edu) received his B.S. degree from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, in 2014. He was a research assistant in the Graduate Institute of Communication Engineering (GICE), NTU in 2016. He is currently studying for his Ph.D. degree in communication theory and systems at the University of California San Diego. His research interests include LTE-Advanced L2/L3 protocol design and next generation wireless communication systems.

Dian-Yu Lin (r04921057@ntu.edu.tw) received his B.S. degree from the Department of Electrical Engineering, National Chiao Tung University, Hsinchu, Taiwan, in 2015. He is currently an M.S. student in GICE, NTU. His research topics include mobile edge computing, fog radio access networks, and cloud radio access networks.

Chia-Fu Lee (noname0225@hotmail.com) graduated from National Tsing Hua University, Hsinchu, Taiwan. He is now studying communication technology at NTU.

Ping-Jung Hsieh (d00921029@ntu.edu.tw) received his B.S. degree in computer science and information engineering from Chang Gung University, Taiwan, in 2009. He received his M.S. degree in communication engineering from National Central University, Taiwan, in 2011. He is currently working toward his Ph.D. degree with the Wireless Mobile Networking Laboratory in the Department of Electrical Engineering, NTU. His research interests include mobile networking (mobility management, new mobility technologies), broadband wireless access networks (5G, 3GPP LTE Advanced, WiMAX), and advanced system architecture aspects (C/U-plane split and mobile edge/fog computing). He has 4G testbed development experience in system/protocol design, evaluation, and implementation.

Hung-Yu Wei (hywei@ntu.edu.tw) received his B.S. degree in electrical engineering from NTU in 1999. He received his M.S. and Ph.D. degrees in electrical engineering from Columbia University in 2001 and 2005, respectively. He was a summer intern at Telcordia Applied Research in 2000 and 2001. He was with NEC Labs America from 2003 to 2005. He joined the Department of Electrical Engineering at NTU in July 2005. He is currently a professor with the Department of Electrical Engineering and GICE at NTU. His research interests include broadband wireless communications, the Internet of Things, and game theoretic models for networking. He actively participates in wireless communications standardization activities, and was a voting member of the IEEE 802.16 Working Group. He was the recipient of the Recruiting Outstanding Young Scholar Award from the Foundation for the Advancement of Outstanding Scholarship in 2006, the K. T. Li Young Researcher Award from ACM's Taipei Chapter and IICM in 2012, the CIEE Excellent Young Engineer Award in 2014, and the NTU Excellent Teaching Award in 2008. He was awarded the Research Project for Excellent Young Scholars from Taiwan's Ministry of Science and Technology in 2014. He was also the recipient of the Wu Ta You Memorial Award from the Ministry of Science and Technology in 2015. Currently, he is the Chair of the IEEE Vehicular Technology Society Taipei Section. He also serves as an Associate Editor of the *IEEE Internet of Things Journal*.

Chun-Ting Chou (chuntingchou@ntu.edu.tw) is an associate professor in GICE, NTU. He received his Ph.D. from the University of Michigan, Ann Arbor in 2004. Before he joined NTU in 2008, he was a senior member of research staff in Philips Research North America, New York, for three years, where he was in charge of standardization for IEEE 802.15.3c, IEEE 802.22, ECMA-368, ECMA-387, and various international standards in wireless communication and networking. His main research interests include dynamic spectrum sharing, medium access control, wireless access for IoT, and 4G/5G mobile networking. He has published more than 30 papers in *IEEE Transactions on Mobile Computing*, *IEEE/ACM Transactions on Networking*, *IEEE Transactions on Wireless Communications*, and *IEEE Transactions on Intelligent Transportation*, IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, and so on. He also holds more than 20 granted patents in wireless system and applications.

Ai-Chun Pang (acpang@csie.ntu.edu.tw) received her B.S., M.S. and Ph.D. degrees in computer science and information engineering from National Chiao Tung University in 1996, 1998, and 2002, respectively. She joined the Department of Computer Science and Information Engineering (CSIE), NTU, in 2002. She is now a professor in CSIE and INM, and is also an adjunct research fellow of the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. Her research interests include the design and analysis of wireless and multimedia networking, mobile communications, and fog/edge computing.

IEEE 5G Learning Series - NJ Edition

Date: Wednesday, April 26, 2017
Venue: The E Hotel Banquet & Conference Center
3050 Woodbridge Avenue, Edison, NJ 08837
http://www.edisonhotelcp.net

TUTORIAL SPEAKERS

- Dr. Sudhir Dixit
  Fellow and Evangelist, Basic Internet Foundation, Oslo, Norway
- Dr. K Raghunandan
  New York Transit
- Yongxing Zhou
  Vice President, Huawei
- Dr. Lingjia Liu
  University of Kansas
- Prof. Jennifer Chen
  Stevens Institute of Technology
- Dr. Jiachen Chen
  Rutgers WINLAB
- Dr. Xiaoxiong (Kevin) Gu
  IBM T.J. Watson Research Center
- Akshay Sharma
  VP at Beesion

Contact Info:
Email: ieee5g-education@ieee.org
http://www.5g.ieee.org/education

IEEE 5G Initiative Learning Series – NJ Edition

5G is not just the next evolution of 4G technology; it is a paradigm shift. 5G is not only evolutionary (providing higher bandwidth and lower latency than current-generation technology); more importantly, 5G is revolutionary, in that it is expected to enable fundamentally new applications with much more stringent requirements in latency (e.g. real time) and bandwidth (e.g. streaming). 5G should help solve the last-mile / last-kilometer problem and provide broadband access to the next billion users on earth at much lower cost because of its use of new spectrum and its improvements in spectral efficiency. 5G is an enabler of exciting use cases that will transform the way people live, work, and engage with their environment. In the short term, 5G can support exciting use cases such as the IoT, smart transportation, eHealth, smart cities and smart homes, industrial automation, and entertainment services.

The IEEE 5G Learning Series is designed to demystify 5G technologies and train technology and industry teams with the knowledge of 5G technologies. This tutorial will provide an understanding of the following topics:

- Introduction to 5G
- 5G RAN
- 5G IOT
- 5G Hardware
- 5G Application
- 5G Core s
- 5G Standards
- 5G Security
- 5G OSS
- 5G Business

Early Registration ends April 6

Attendee Registration: https://events.vtools.ieee.org/m/44054

Patron/Exhibitor Registration:
https://meetings.vtools.ieee.org/meeting_view/list_meeting/44174

Dr. Amruthur Narasimhan, IEEE 5G NJ Tutorial Chair, anarasimhan@ieee.org
Dr. Ashutosh Dutta, IEEE 5G Initiative Co-Chair, ashutosh.dutta@ieee.org
Dr. Rulei Ting, IEEE 5G Education working group Co-Chair, rt@ieee.org

Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges

Tuyen X. Tran, Abolfazl Hajisami, Parul Pandey, and Dario Pompili

The authors envision a real-time, context-aware collaboration framework that lies at the edge of the RAN, comprising MEC servers and mobile devices, and amalgamates the heterogeneous resources at the edge. Specifically, they introduce and study three representative use cases ranging from mobile edge orchestration, collaborative caching and processing, and multi-layer interference cancellation.

ABSTRACT

MEC is an emerging paradigm that provides computing, storage, and networking resources within the edge of the mobile RAN. MEC servers are deployed on a generic computing platform within the RAN, and allow for delay-sensitive and context-aware applications to be executed in close proximity to end users. This paradigm alleviates the backhaul and core network and is crucial for enabling low-latency, high-bandwidth, and agile mobile services. This article envisions a real-time, context-aware collaboration framework that lies at the edge of the RAN, comprising MEC servers and mobile devices, and amalgamates the heterogeneous resources at the edge. Specifically, we introduce and study three representative use cases ranging from mobile edge orchestration, collaborative caching and processing, and multi-layer interference cancellation. We demonstrate the promising benefits of the proposed approaches in facilitating the evolution to 5G networks. Finally, we discuss the key technical challenges and open research issues that need to be addressed in order to efficiently integrate MEC into the 5G ecosystem.

INTRODUCTION

Over the last few years, our daily lifestyle is increasingly exposed to a plethora of mobile applications for entertainment, business, education, health care, social networking, and so on. At the same time, mobile data traffic is predicted to continue doubling each year. To keep up with these surging demands, network operators have to spend enormous efforts to improve users' experience while maintaining healthy revenue growth. To overcome the limitations of current radio access networks (RANs), two emerging paradigms have been proposed:

- Cloud RAN (C-RAN), which aims at the centralization of base station (BS) functions via virtualization
- Mobile edge computing (MEC), which proposes to empower the network edge

While the two technologies propose to move computing capabilities in a different direction (to the cloud vs. to the edge), they are complementary, and each has a unique position in the fifth generation (5G) ecosystem.

As depicted in Fig. 1, MEC servers are implemented directly at the BSs using a generic computing platform, allowing the execution of applications in close proximity to end users. With this position, MEC can help fulfill the stringent low-latency requirement of 5G networks. Additionally, MEC offers various network improvements, including:

- Optimization of mobile resources by hosting compute-intensive applications at the network edge
- Pre-processing of large data before sending it (or some extracted features) to the cloud
- Context-aware services with the help of RAN information such as cell load, user location, and allocated bandwidth

Although the MEC principle also aligns with the concept of *fog computing* [1], and the two are often referred to interchangeably, they slightly differ from each other. While fog computing is a general term that opposes cloud computing in bringing the processing and storage resources to the lower layers, MEC specifically aims at extending these capabilities to the edge of the RAN with new function splitting and a new interface between the BSs and the upper layer. Fog computing is most commonly seen in enterprise-owned gateway devices, whereas MEC infrastructure is implemented and owned by the network operators.

Fueled by the potential capabilities of MEC, we propose a real-time context-aware collaboration framework that lies at the edge of the cellular network and works side by side with the underlying communication network. In particular, we aim at exploring the synergies among connected entities in the MEC network to form a heterogeneous computing and storage resource pool. To illustrate the benefits and applicability of MEC collaboration in 5G networks, we present three use cases including mobile edge orchestration, collaborative video caching and processing, and multi-layer interference cancellation. These initial target scenarios can be used as the basis for the formulation of a number of specific applications.

The remainder of this article is organized as follows. In the following section, we present the state of the art on MEC. Then we provide a comparison between MEC and C-RAN in various

Digital Object Identifier:
10.1109/MCOM.2017.1600863

The authors are with Rutgers University.

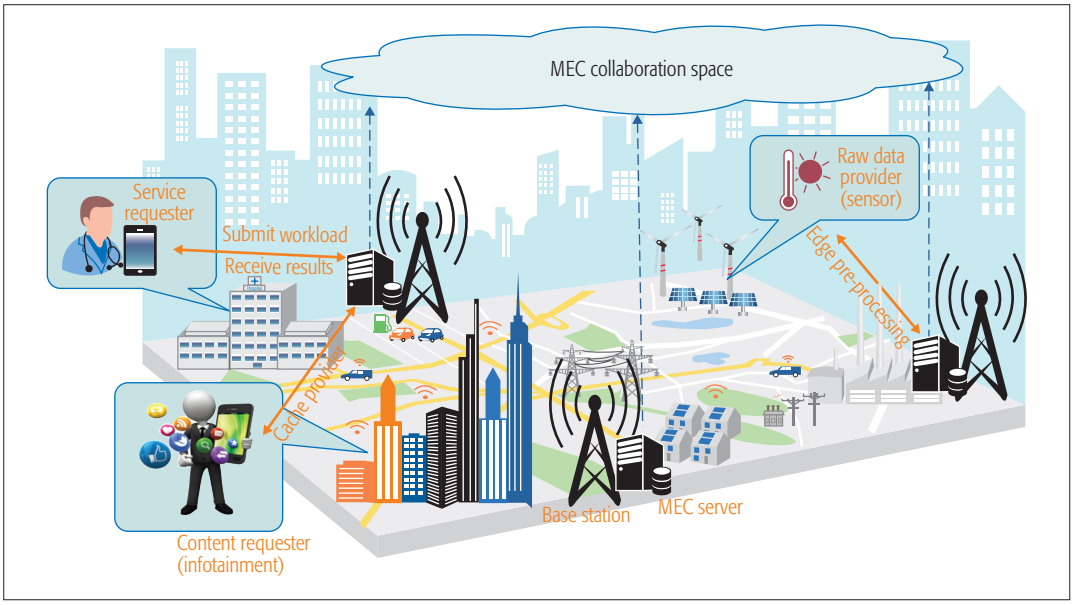


Figure 1. Illustration of Mobile Edge Computing Network.

features. Following that, we describe the three case studies to illustrate the applicability and benefits of the collaborative MEC paradigm. Then we highlight some key challenges and open research issues that need to be tackled. Finally, we draw our conclusions in the final section.

State of the Art

In 2013, Nokia Networks introduced a very first real-world MEC platform [2] in which the computing platform — radio applications cloud servers (RACS) — is fully integrated with the Flexi Multiradio base station. Under a scenario of a "smarter city" [3], IBM discusses how operators can leverage the capabilities of mobile edge network virtualization to deploy disruptive services for consumers and enterprises. Saguna also introduces their fully virtualized MEC platform, Open-RAN [4], which can provide an open environment for running third-party MEC applications. Recently, the European Telecommunications Standards Institute (ETSI) formed a MEC Industry Specifications Group (ISG) in order to standardize and moderate the adoption of MEC within the RAN [5].

From the theoretical perspective, the authors in [6] consider the computation offloading problem in a multi-cell MEC network, where a dense deployment of radio access points facilitates proximity high-bandwidth access to computational resources but also increases inter-cell interference. The authors in [7] provide a collective overview of the opportunities and challenges of "fog computing" in the networking context of the Internet of Things (IoT). Several case studies are presented to highlight the potential and challenges of the fog control plane such as interference, control, configuration, and management of networks, and so on (http://Fogresearch.org).

In summary, prior works on MEC focused on feasibility of MEC-RAN integration, deployment scenarios, and potential services and applications. *In contrast to existing works on MEC, which do not explore the synergies among the MEC entities, this article takes one step further by proposing a collaborative MEC paradigm and presents three strong use cases to efficiently leverage this collaboration space.*

MEC vs. C-RAN

A redesigned centralization of RAN is proposed as C-RAN, where the physical layer communication functionalities are decoupled from the distributed BSs and are consolidated in a virtualized central processing center. With its centralized nature, it can be leveraged to address the capacity fluctuation problem and to increase system energy efficiency in mobile networks [8]. Besides an approach to 5G standardization, C-RAN can provide new opportunities for IoT, opening up a new horizon of ubiquitous sensing, interconnection of devices, service sharing, and provisioning to support better communication and collaboration among people and things in a more distributed and dynamic manner. The integration of cloud provider, edge gateways, and end devices can support powerful processing and storage facilities to massive IoT data streams (big data) beyond the capability of individual "things" as well as provide automated decision making in real time. Thus, the C-RAN and IoT convergence can enable the development of new innovative applications in various emerging areas such as smart cities, smart grids, smart healthcare, and others aimed at improving all aspects of human life.

The full centralization principle of C-RAN, however, entails the exchange of radio signals between the radio heads and cloud processing unit, which imposes stringent requirement to the fronthaul connections in terms of throughput and latency. On the other hand, the MEC paradigm is useful in reducing latency and improving localized user experience, but the amount of processing power and storage is orders of magnitude below that of the centralized cloud in C-RAN. In Table 1, we summarize the comparison between MEC and C-RAN in various aspects. One important note is that MEC does not contradict with C-RANs but rather complement them. For example, an appli-

In contrast to existing works on MEC, which do not explore the synergies among the MEC entities, this article takes one step further by proposing a collaborative MEC paradigm and presents three strong use cases to efficiently leverage this collaboration space.

The proposed novel resource management framework lies at the intermediate edge layer and orchestrates both the horizontal collaboration at the end-user layer and the MEC layer as well as the vertical collaboration between end users, edge nodes, and cloud nodes.

	MEC	C-RAN
Location	Co-located with base stations or aggregation points.	Centralized, remote data centers.
Deployment planning	Minimal planning with possible ad hoc deployments.	Sophisticated.
Hardware	Small, heterogeneous nodes with moderate computing resources.	Highly capable computing servers.
Fronthaul requirements	Fronthaul network bandwidth requirements grow with the total amount of data that need to be sent to the core network after being filtered/processed by MEC servers.	Fronthaul network bandwidth requirements grow with the total aggregated amount of data generated by all users.
Scalability	High	Average, mostly due to expensive fronthaul deployment.
Application delay	Support time-critical applications that require latencies less than tens of milliseconds.	Support applications that can tolerate round-trip delays on the order of a few seconds or longer.
Location awareness	Yes	N/A
Real-time mobility	Yes	N/A

Table 1. Comparison of features: MEC vs. C-RAN.

cation that needs to support very low end-to-end delay can have one component running in the MEC cloud and other components running in the distant cloud.

In the following sections, we present our case studies where we propose novel scenarios and techniques to take advantage of the collaborative MEC systems.

Case Study I: Mobile Edge Orchestration

In spite of the limited resources (e.g., battery, CPU, memory) on mobile devices, many computation-intensive applications from various domains such as computer vision, machine learning, and artificial intelligence are expected to work seamlessly with *real-time* responses. However, the traditional way of offloading computation to the remote cloud often leads to unacceptable delay (e.g., hundreds of milliseconds [9]) and heavy backhaul usage. Due to its distributed computing environment, MEC can be leveraged to deploy applications and services as well as to store and process content in close proximity to mobile users. This would enable applications to be split into small tasks with some of the tasks performed at the local or regional clouds as long as the latency and accuracy are preserved.

In this case study, we envision a collaborative distributed computing framework where resource-constrained end-user devices outsource their computation to the upper-layer computing resources at the edge and cloud layers. Our framework extends the standard MEC originally formulated by ETSI, which only focuses on individual MEC entities and on the vertical interaction between end users and a single MEC node. Conversely, our proposed collaborative framework will bring many individual entities and infrastructures to collaborate with each other in a distributed system. In particular, our framework oversees a hierarchical architecture consisting of:

- *End user*, which implies both mobile and static end-user devices such as smartphones, sensors, and actuators
- *Edge nodes*, which are the MEC servers co-located with the BSs
- *Cloud node*, which is the traditional cloud-computing server in a remote data center

Our novel resource management framework lies at the intermediate edge layer and orchestrates both the *horizontal* collaboration at the end-user layer and the MEC layer as well as the *vertical* collaboration between end users, edge nodes, and cloud nodes. The framework will make dynamic decisions on "what" and "where" the tasks in an application should be executed based on the execution deadline, network conditions, and device battery capacity.

There have been a number of works in the mobile computing domain where data from the local device is uploaded to the cloud for further processing [10] or executed locally via approximate computing [11] to combat the problem of limited resources. In [12] we focused on the "extreme" scenario in which the resource pool was composed purely of proximal mobile devices. In contrast, MEC introduces a new stage of processing such that the edge nodes can analyze the data from nearby end users and notify the cloud node for further processing only when there is a significant change in data or accuracy of results. In addition, sending raw sensor values from end users to the edge layer can overwhelm the fronthaul links; hence, depending on the storage and compute capabilities of user devices and the network conditions, the MEC orchestrator can direct the end users to extract features from the raw data before sending to the edge nodes.

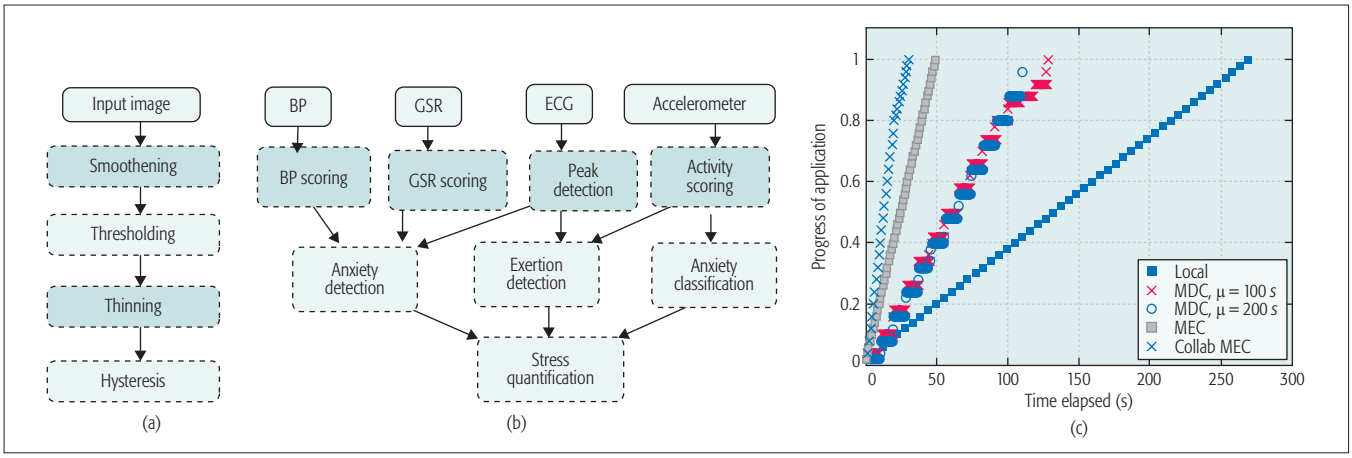


Figure 2. Block diagram showing tasks of different mobile applications in a) image processing domain (canny edge detection); b) ubiquitous health-care domain (stress quantification). The darker blocks in these applications represent the computationally intensive tasks of the applications that can be offloaded to the remote resources (edge and cloud); c) comparison of different startegies to execute computationally intensive mobile applications.

In Figs. 2a and 2b, we illustrate two mobile applications from different domains that are good candidates for being executed at the edge. The darker blocks in these applications represent the computation-intensive tasks of the applications that can be offloaded to the upper-level resources (edge and cloud). In Fig. 2c we compare the time taken for execution of the mobile application represented in Fig. 2a (canny edge detection) by using different strategies:

- Executing the application locally on the mobile device (Local)
- Distributing tasks to proximal mobile devices forming a mobile device cloud (MDC) [12]
- Offloading the tasks to a single MEC server (MEC)
- To two collaborating MEC servers (collab MEC), respectively.

For execution in an MDC we model the mobility patterns of devices in the proximity as a normal distribution with mean availability duration of devices varying with $\mu = \{100, 200\}$ s and $\sigma = 5$ s. We assume that the local mobile devices connect with the MEC server on a 1 Mb/s link. The mobile devices involved in the experiment include two Samsung Galaxy Tabs and four smartphones (two ZTE Avid N9120s and two Huawei M931s). For MEC servers we used two desktops with Intel Core i7 CPU at 3.40 GHz and 16 GB RAM. We execute the application in Fig. 2a by using input data from the Berkeley image segmentation and benchmark dataset. Resolution of each image is 481 × 321 pixels. A task consists of finding edges of 20 images from the dataset. For the current simulation, we use a round-robin technique for the MDC where all the devices are given equal tasks. Sophisticated task allocation algorithms can be run at the arbitrator to decide how many tasks to run at each service provider based on the computational capabilities of different service providers. After execution of the tasks, the service provider returns the task to the service requester. In Fig. 2c we see that the performance of execution on a single MEC server is significantly better than the execution on a local device and MDC. The gain in terms of execution time on using collaborative MEC over execution of the application on a single MEC server is around 40 percent.

The example above illustrates the benefit of the collaborative MEC framework in reducing execution time of the two image processing tasks. The extension of this strategy will greatly benefit the service requesters, which are health analytics providers in this case, as they see lower latency in the execution of the application as the MEC servers are at the BS rather than at the cloud. These service requesters require processing of large data, and the MEC servers expedite the processing time by dividing the processing between MEC servers (extracting features from the raw data) and cloud resources (running computation-intensive applications using extracted features as input data). This leads to faster availability results for the data analytics expert and also gives faster results to patients requesting results.

Currently, to present preliminary results we use a simple image processing application. However, we believe that a compute-intensive application (e.g., real-time activity detection with significant variations in execution time of tasks) or a data-intensive application (e.g., real-time face detection in a video with a large volume of input data) will require a powerful computing environment like ours to make dynamic decisions on what and where are the tasks to be executed based on real-time conditions, which will make application execution via collaborative MEC even more challenging.

Case Study II: Collaborative Video Caching and Processing

Mobile video streaming traffic is predicted to account for 72 percent of the overall mobile data traffic by 2019 [13], posing immense pressure on network operators. To overcome this challenge, edge caching has been recognized as a promising solution, by which popular videos are cached in the BSs or access points so that demands from users for the same content can be accommodated easily without duplicate transmission from remote servers. This approach helps substantially reduce backhaul usage and content access delay. While

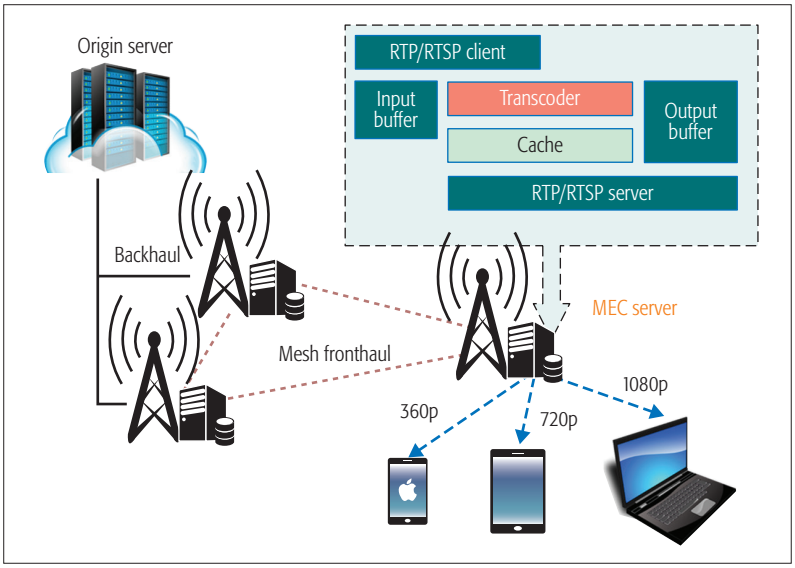


Figure 3. Illustration of collaborative video caching and processing framework deployed on an MEC network.

content caching and delivery techniques in wireless networks have been studied extensively (e.g., [14, references therein]), existing approaches rarely exploit the synergy of caching and computing at the cache nodes. Due to the limited cache storage at individual BSs, the cache hit rate is still moderate. Several solutions have considered collaborative caching, in which a video request can be served using not only the local BS's cache, but also the cached copy at neighboring BSs via the backhaul links [15].

With the emergence of MEC, it is possible to not only perform edge caching but also edge processing. Our approach will leverage edge processing capability to improve caching performance/efficiency. Such a joint caching and processing solution will trade off storage and computing resources with backhaul bandwidth consumption, which directly translates into sizable network cost saving. Due to the heterogeneity of users' processing capabilities and the variance of network connections, user preference and demand toward a specific video might be different. For example, users with highly capable devices and fast network connections usually prefer high-resolution videos, whereas users with low processing capabilities or low bandwidth connections may not enjoy high-quality videos because the delay is large and the video may not fit within the device's display. Leveraging such behavior, adaptive bit rate (ABR — https://en.wikipedia.org/wiki/Adaptive_bitrate_streaming) streaming techniques have been developed to improve the quality of delivered video on the Internet as well as wireless networks. Examples of such techniques include Apple HTTP Live Streaming (HLS), Microsoft Smooth Streaming, and Adobe Systems HTTP Dynamic Streaming. In ABR streaming, the quality of the streaming video is adjusted according to the user device's capabilities, network connection, and specific request. Existing video caching systems often treat each request for a video version equally and independently, without considering their transcoding relationship, resulting in moderate benefits.

In this case study, we exploit both ABR streaming and collaborative caching to improve the caching benefits beyond what can be achieved by traditional approaches. The proposed collaborative video caching and processing framework deployed on a MEC network [16] is illustrated in Fig. 3. Given the storage and computing capabilities, each MEC server acts as a cache server as well as a transcoding server. These servers collaborate with each other to not only provide the requested video but also transcode it to an appropriate variant. Each variant is a bit rate version of the video, and a higher bit rate version can be transcoded into lower bit rate ones. The potential benefits of this strategy are three-fold:

- The content origin servers need not generate all variants of the same video.
- Users with various capabilities and network conditions will receive videos that are suited for their capabilities, as content adaptation is more appropriately done at the network edge.
- Collaboration among the MEC servers enhances cache hit ratio and balance processing load in the network.

In our proposed joint collaborative caching and processing strategy, referred to as *CoPro-CoCache*, we distribute the most popular videos in the serving cell of each BS to the corresponding cache server of that BS until the cache storage is full. When a user requests a video that requires transcoding from a different version in the cache, the transcoding task is assigned to the MEC server having lower load, which could be the MEC server storing the original video version (data provider node) or the serving MEC server (delivery node). This helps balance the processing load in the network.

To illustrate the potential benefits of the proposed approach, we perform numerical simulation on a representative RAN with five BSs, each equipped with a MEC server that performs caching and transcoding. We assume a library of 1000 videos is available for download. The video popularity requested at each BS follows a Zipf distribution with parameter 0.8, that is, the probability that an incoming request is for the ith most popular video is proportional to $1/i^{0.8}$. In order to obtain a scenario where the same video can have different popularities at different locations, we randomly shuffle the distributions at different BSs. Video request arrival follows a Poisson distribution with same rate at each BS. In Figs. 4a and 4b we compare the performance of four caching strategies in terms of backhaul traffic reduction. It can be seen that utilizing processing capabilities significantly helps reduce the backhaul traffic load. In addition, our proposed *CoPro-CoCache* strategy explores the synergies of processing capabilities among the MEC servers, rendering additional performance gain. Figure 4c illustrates the processing resource utilization of the *CoPro-CoCache* scheme vs. different video request arrival rates and cache capacity. We observe that the processing utilization increases with arrival rate and moderate cache capacity; however, it decreases at high cache capacity. This is because with high cache capacity, we can store almost all the popular videos and their variants, and thus there are fewer requests requiring transcoding.

While choosing the optimal bit rate for video streaming can enhance instant download throughput, existing client-based bit rate selection may not be able to adapt fast enough to the rapidly

varying conditions, leading to underutilization of radio resources and suboptimal user experience. A promising solution is to use a RAN analytic agent at the MEC server to inform the video server of the optimal bit rate to use given the radio conditions for a particular video request from an end user. Designing an efficient solution to address bit rate adaption with respect to channel conditions is still an open problem.

Case Study III: Two-Layer Interference Cancellation

Deploying more small cell BSs can improve spectral efficiency in cellular networks, however, making inter-cell interference become more prominent. To mitigate such interference, a promising approach is to employ coordinated multipoint (CoMP) transmission and reception techniques. In CoMP, a set of neighboring cells are divided into clusters; within each cluster, the BSs are connected to each other via a fixed backhaul processing unit (BPU) and exchange channel state information (CSI) as well as mobile station (MS) signals to cancel the intra-cluster interference. However, CoMP does not take into account the inter-cluster interference, resulting in moderate improvement in system capacity. Furthermore, the additional processing required for multi-site reception/transmission, CSI acquisition, and signaling exchanges among different BSs could add considerable delay and thus limit the cluster size in order to comply with the stringent delay requirement in 5G networks. In addition, applying CoMP for all users might be unnecessary as certain users, especially those at the cell centers, often have high levels of signal-to-interference-plus-noise ratio (SINR) and do not cause intense interference to the neighboring BSs.

To overcome the existing challenges of CoMP, and reduce the latency and bandwidth between the BSs and the BPU, we advocate a two-layer interference cancellation strategy for an uplink MEC-assisted RAN. In particular, based on the channel quality indicator (CQI) of each user, our solution identifies "where" to process its uplink signal so as to reduce complexity, delay, and bandwidth usage. In a MEC-assisted RAN, we have access to the computational processing at the BSs, and the signal demodulation of the cell center MSs can be done in local BSs (layer 1). This means that the system performance for cell center MSs relies on a simple single transmitter and receiver. On the other hand, since the SINRs of cell edge MSs are often low, their signals should be transmitted to the BPU (layer 2) for further processing. In this case, the BPU has access to all the celledge MSs from different cells and is able to improve their SINRs via coordinated processing.

As illustrated in Fig. 5, each red dotted circle indicates the interference region of the corresponding cell, which is defined as a region within which if MSs from other cells moved in, they could render "intense" interference at the BS serving the cell. Since MS #1 is a cell center MS and is outside the interference region of BSs #2 and #3, its interference at those BSs is low due to the high path loss; hence, there is no need to employ coordinated interference cancellation for MS #1, and thus its signal demodulation can be performed at the edge layer. Conversely, since MS #2 is a cell

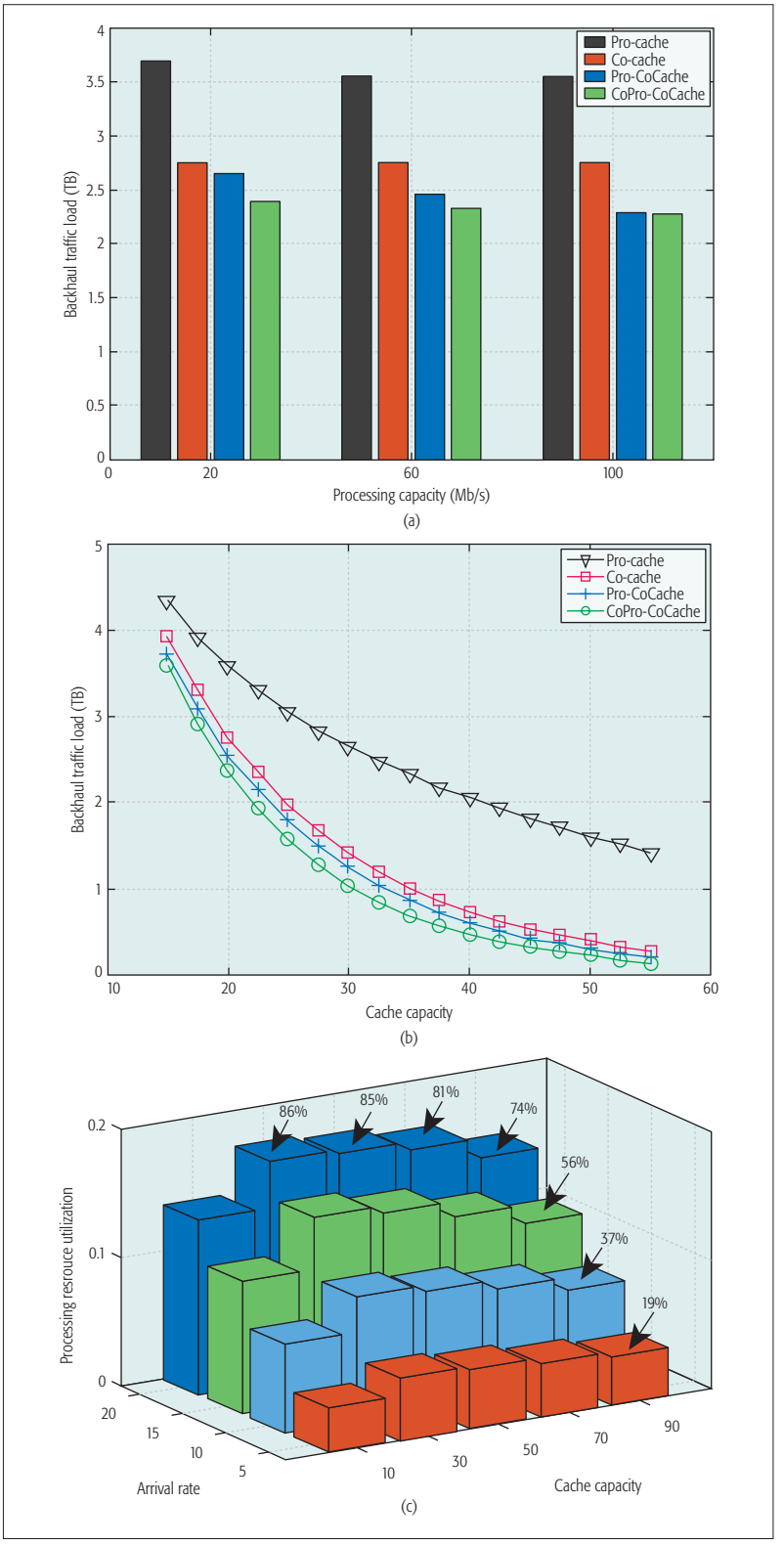


Figure 4. Considered caching strategies: Pro-Cache—non-collaborative caching with processing; Co-Cache—collaborative caching without processing; Pro-CoCache—collaborative caching with processing; and CoPro-Co-Cache—collaborative caching with collaborative processing (proposed). Video duration is set to 10 min, and each video has four variants with relative bit rates of 0.82, 0.67, 0.55, and 0.45 of the original video bit rate (2 Mb/s): a) backhaul traffic load vs. processing capacity (Mb/s) with cache capacity = 30 percent library size; b) backhaul traffic load vs. cache capacity; c) processing resource utilization vs. arrival rate (request/BS/min) and cache capacity. In b and c, we set processing capacity = 40 Mb/s.

Mobile-Edge Computing enables a capillary distribution of cloud computing capabilities to the edge of the radio access network. This emerging paradigm allows for execution of delay-sensitive and context-aware applications in close proximity to the end-users while alleviating backhaul utilization and computation at the core network.

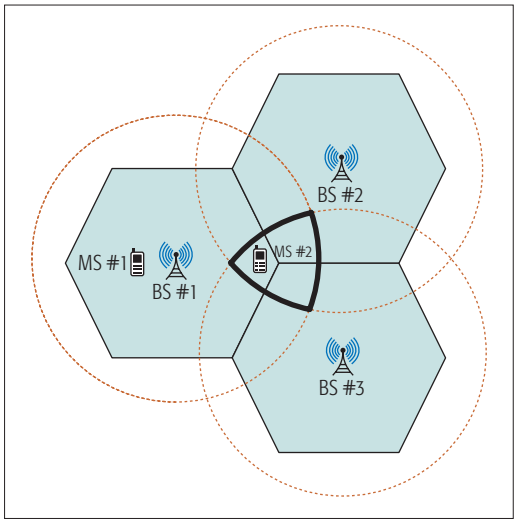


Figure 5. MSs #1 and #2 are located at cell center and cell edge regions, respectively. Since MS #1 is far from the neighboring BSs, signal demodulation can be performed at the edge (layer 1). However, MS #2 is located at the cell edge region, and its interference to the neighbouring BSs should be canceled at the upper layer (layer 2).

edge MS and is located in the interference region of BSs #2 and #3, there may be an intense interference from MS #2 to BSs #2 and #3; thus, coordinated interference cancellation at the upper layer is needed to cancel this interference, and the BS should transmit the raw data to the upper layer for further processing.

Challenges and Open Research Issues

The decentralization of cloud computing infrastructure to the edge brings various benefits that contribute to the 5G evolution, and at the same time introduces new challenges and open research issues, highlighted in the following.

Resource Management: The computing and storage resources in an individual MEC platform are expected to be limited and may be able to support a constrained number of applications with moderate needs for such resources. Currently, network providers often race for extensively standalone infrastructures to keep up with the demand while struggling with lower return on investment. An alternative approach such as MEC as a service may need to be considered, whereby operators' resources can be opened up for interested service providers to request or relinquish based on service demand.

Interoperability: MEC infrastructures owned by different network providers should be able to collaborate with each other as well. This necessitates the specification of common collaboration protocols, also allowing for service providers to access network and context information regardless of their deployment place.

Service Discovery: Exploiting the synergies of distributed resources and various entities, as envisioned in our mobile edge orchestration framework, requires discovery mechanisms to find appropriate nodes that can be leveraged in a decentralized setup. Automatic monitoring of the heterogeneous resources and accurate synchronization across multiple devices are also of great importance.

Mobility Support: In a small cell network, the range of each individual cell is limited. Mobility support becomes more important, and a solution for fast process migration may become necessary.

Fairness: Ensuring fair resource sharing and load balancing is also an essential problem. There is potential that a small number of nodes could carry the burden of processing, while a large number of nodes would contribute little to the efficiency of the distributed network.

Security: Security issues might hinder the success of the MEC paradigm if not carefully considered. Existing centralized authentication protocols might not be applicable for some parts of the infrastructure that have limited connectivity to the central authentication server. It is also important to implement trust management systems that are able to exchange compatible trust information with each other, even if they belong to different trust domains. Furthermore, as service providers want to acquire user information to tailor their services (e.g., content providers want to know users' preferences and mobility patterns to proactively cache their contents, as discussed in case study II), there is a great challenge to the development of privacy protection mechanisms that can efficiently protect users' location and service usage.

Conclusions

Mobile edge computing enables a capillary distribution of cloud computing capabilities to the edge of the radio access network. This emerging paradigm allows for execution of delay-sensitive and context-aware applications in close proximity to end users while alleviating backhaul utilization and computation at the core network. This article proposes to explore the synergies among connected entities in the MEC network to form a heterogeneous resource pool. We present three representative use cases to illustrate the benefits of MEC collaboration in 5G networks. Technical challenges and open research issues are highlighted to give a glimpse of the development and standardization roadmap of the mobile edge ecosystem.

Acknowledgment

This work was supported in part by the National Science Foundation (NSF) under Grant No. CNS-1319945.

References

[1] F. Bonomi *et al.*, "Fog Computing and Its Role in the Internet of Things," *Proc. 1st ACM Wksp. Mobile Cloud Computing*, 2012, pp. 13–16.
[2] Intel and Nokia Siemens Networks, "Increasing Mobile Operators' Value Proposition with Edge Computing," technical brief, 2013.
[3] IBM, "Smarter Wireless Networks; Add Intelligence to the Mobile Network Edge," Thought Leadership white paper, 2013.
[4] Saguna and Intel, "Using Mobile Edge Computing to Improve Mobile Network Performance and Profitability," white paper, 2016.
[5] Y. C. Hu *et al.*, "Mobile Edge Computing — A Key Technology Towards 5G," ETSI white paper, vol. 11, 2015.
[6] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," *IEEE Trans. Signal Info. Processing over Networks*, vol. 1, no. 2, 2011, pp. 89–1035.
[7] M. Chiang and T. Zhang, "Fog and IoT: An Overview of Research Opportunities," *IEEE Internet of Things J.*, vol. 3, no. 6, 2016, pp. 854–64.

[8] D. Pompili, A. Hajisami, and T. X. Tran, "Elastic Resource Utilization Framework for High Capacity and Energy Efficiency in Cloud RAN," *IEEE Commun. Mag.*, vol. 54, no. 1, Jan. 2016, pp. 26–32.
[9] E. Cuervo *et al.*, "MAUI: Making Smartphones Last Longer with Code Offload," *Proc. ACM Int'l. Conf. Mobile Systems Applications Services*, 2010, pp. 49–62.
[10] M. S. Gordon *et al.*, "COMET: Code Offload by Migrating Execution Transparently," *Proc. USENIX Conf. Operating Systems Design and Implementation*, Oct. 2012, pp. 93–106.
[11] P. Pandey and D. Pompili, "Exploiting the Untapped Potential of Mobile Distributed Computing via Approximation," *Pervasive and Mobile Computing*, 2017.
[12] H. Viswanathan, P. Pandey, and D. Pompili, "Maestro: Orchestrating Concurrent Application Workflows in Mobile Device Clouds," *Proc. Wksp. Distrib. Adaptive Systems, Int'l. Conf. Autonomic Computing*, July 2016, pp. 257–62.
[13] Global Mobile Data Traffic Forecast, Update 2014–2019, White Paper c11-520862, Cisco Visual Networking Index.
[14] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 82–89.
[15] T. X. Tran and D. Pompili, "Octopus: A Cooperative Hierarchical Caching Strategy for Cloud Radio Access Networks," *Proc. IEEE Int'l. Conf. Mobile Ad Hoc Sensor Systems*, Oct. 2016, pp. 154–62.
[16] T. X. Tran *et al.*, "Collaborative Multi-bitrate Video Caching and Processing in Mobile-Edge Computing Networks," *Proc. IEEE Annual Conf. Wireless On-demand Network Systems Services*, Feb. 2017.

Biographies

Tuyen X. Tran (tuyen.tran@cac.rutgers.edu) is working toward his Ph.D. degree in electrical and computer engineering (ECE) at Rutgers University under the guidance of Dr. Pompili. He received his M.Sc. degree in ECE from the University of Akron, Ohio, in 2013, and his B.Eng. degree (Honors Program) in electronics and telecommunications from Hanoi University of Technology, Vietnam, in 2011. His research interests are in the application of optimization, statistics, and game theory to wireless communications and cloud computing.

Abolfazl Hajisami (hajisamik@cac.rutgers.edu) started his Ph.D. program in ECE at Rutgers University in 2012. Currently, he is pursuing research in the fields of C-RAN, cellular networking, and mobility management under the guidance of Dr. Pompili. Previously, he received his M.S. and B.S. from Sharif University of Technology and Shahid Beheshti University, Tehran, Iran, in 2010 and 2008, respectively. His research interests are wireless communications, cloud radio access networks, statistical signal processing, and image processing.

Parul Pandey (parul_pandey@cac.rutgers.edu) is a Ph.D. candidate in the Department of ECE at Rutgers University. She is currently working on mobile and approximate computing, cloud-assisted robotics, and underwater acoustic communications under the guidance of Dr. Pompili as a member of the Cyber-Physical Systems Laboratory (CPS-Lab). Previously, she received her B.S. degree in electronics and communication engineering from Indira Gandhi Institute of Technology, Delhi, India, and her M.S. degree in ECE from the Univeristy of Utah in 2008 and 2011, respectively.

Dario Pompili [SM] (pompilig@cac.rutgers.edu) is an associate professor with the Department of ECE at Rutgers University, where he directs the CPS-Lab. He received his Ph.D. in ECE from the Georgia Institute of Technology in 2007. He previously received his Laurea (combined B.S. and M.S.) and doctorate degrees in telecommunications and systems engineering from the University of Rome "La Sapienza," Italy, in 2001 and 2004, respectively. He is a recipient of the NSF CAREER '11, ONR Young Investigator Program '12, and DARPA Young Faculty '12 awards. He is a Senior Member of the ACM.

SDN USE CASES FOR SERVICE PROVIDER NETWORKS: PART 2


Ashwin Gumaste


Vishal Sharma


Deepak Kakadia


Jennifer Yates


Axel Clauberg


Mirko Voltolini

Software defined networking (SDN) is being heralded as a major breakthrough for provisioning new services, commoditizing network hardware and reducing service provider capital and operational expenditures. Decoupling the control plane from the data plane has its inherent advantages in terms of offering flexibility and presenting new service capabilities, but also presents operational and functional challenges. Such decoupling would create a new set of use cases, and equipment and network best practices that providers would have to imbibe in the near future. The key scope of this Feature Topic (FT) is to understand, from an implementation perspective, how SDN can be implemented in very large networks using commodity gear or white boxes through use cases.

Many service providers have announced early deployments of SDN platforms. Network equipment vendors, in parallel, have announced the "SDNizing" of their equipment portfolio to meet the emerging needs of service providers. Apart from the initial euphoria surrounding SDN, there is a need for the technical community and the ecosystem to understand the set of services that a large provider would offer using SDN. Similarly, it is of interest to understand how new SDN-centric platforms by vendors would differ from those offered earlier, and whether just pushing the functionality to a controller is the key to success, or an overall architecture that incorporates such an SDN-centric platform must also be devised. Migration policies from a standard telco network to one that supports SDN is the key to understanding how SDN can be successful. In addition, the role of SDN within specific use case environments such as data centers, wireless backhaul and fronthaul, and large enterprises or application-specific providers is important to note.

In this regard, the FT invited contributions and received 50 papers from across four continents. Due to the high volume of papers, the FT was distributed into two issues — one in October 2016 and the other in April 2017. This second part presents seven articles covering a wide gamut of ideas:

The first article, by A. Rostami *et al.*, "Orchestration of RAN and Transport Networks Resources for 5G: An SDN Approach," discusses how SDNs can bring programmability in the transport and radio domains, and discusses use cases for SDN in 5G networks.

The second article, by P. Ventre *et al.*, "SDN-Based IP and Layer 2 Services with an Open Networking Operating System in the GÉANT Service Provider Network," discusses the GÉANT project, particularly focusing on compatibility with BGP and ONOS for a 500 Gb/s continental provider network that uses SDN for research and educational networking.

The third article, by G. Biczok *et al.*, "Manufactured by Software: SDN-Enabled Multi-Operator Composite Services with the 5G Exchange" introduces the 5G Exchange (5GEx) concept that builds on SDN and NFV, and facilitates the provisioning of multi-operator 5G services by means of inter-operator management and orchestration of virtualized network, compute, and storage resources.

The fourth article, by A. Rodriguez-Natal *et al.*, "Global State, Local Decisions: Decentralized NFV for ISPs via Enhanced SDN," discusses the advantages of a decentralized NFV architecture supplemented by an SDN infrastructure in provider networks. Bottleneck components for migration to NFV are quantitatively analyzed.

The fifth article, by M. Menth *et al.*, "Resilient Integration of Distributed High-Performance Zones into the BelWue Network Using OpenFlow," discussed the design philosophy behind the BelWue network that interconnects universities in Germany's high performance network zones, using OpenFlow and BGP. The authors illustrate a prototype and discuss field trial results.

The sixth article, by D. Suh *et al.*, "Toward Highly Available and Scalable Software Defined Networks for Service Provider Networks," analyzes two well-known open source projects, OpenDaylight (ODL) and Open Network Operating System (ONOS), from the perspective of high availability (i.e., network state database replication/synchronization and controller failover mechanisms) and scalability (i.e., network state database partition/distribution and controller assignment mechanisms).

The seventh article, by A. Heurtas *et al.*, "Enabling Highly Dynamic Mobile Scenarios with Software Defined Network-

ing," presents a mobility-aware and policy-based on-demand control network solution oriented to the SDN paradigm.

BIOGRAPHIES

ASHWIN GUMASTE (ashwin@ashwin.name) is currently the Institute Chair Associate Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology (IIT) Bombay. He was a visiting scientist with the Massachusetts Institute of Technology, Cambridge, in the Research Laboratory for Electronics from 2008 to 2010. He was previously with Fujitsu in the Photonics Networking Laboratory (2001–2005) and prior to that with Cisco Systems. His work on light-trails has been widely referred to, deployed, and recognized by both industry and academia. His recent work on omnipresent Ethernet has been adopted by tier-1 service providers and also resulted in the largest ever acquisition between any IIT and the industry. This has led to a family of transport products. He has 23 granted U.S. patents and has published about 150 papers in refereed conferences and journals. He has also authored three books on broadband networks. Due to his many research achievements and contributions, he was awarded the Swaranajayanti Fellowship, India's highest scientific award below the age of 40; the Government of India's DAE-SRC Outstanding Research Investigator Award in 2010; the Department of Space Vikram Sarabhai award in 2012; the IBM Faculty award, as well as the Indian National Academy of Engineering's (INAE) Young Engineer Award (2010).

VISHAL SHARMA [F] is an international telecom expert who has executed assignments for clients on four continents, and worked both in the United States and India. He is an innovator, researcher, engineer, and entrepreneur rolled into one. His multi-faceted roles over the last 20+ years have included entrepreneur, technologist, technical leader, academic (professor of electrical engineering), educator, technology evangelist and advisor, technology writer, editor, and speaker, and researcher in academia and in industry labs. He is currently principal at Metanoia, Inc., a niche Bay Area firm providing expertise to leaders in telecom. His expertise spans core IP network design (MPLS, traffic engineering, VPNs), metro and access network architectures, backhaul design, and both wireline and wireless (4G/LTE, WiMAX) technologies. His current focus is on techniques to optimize a carrier's strategy, business models, operations, and networks. He has concentrated on the application of SDN, NFV, and cloud computing to streamline carrier operations, and on the impact of ubiquitous IoT on network and protocol design. He has also applied this to 4G/LTE and emerging 5G networks. He was on the Advisory Board for the Software-Defined Infrastructure & Cloud Infrastructure tracks for TiECon '13 and TiECon '14 (the world's largest conference on technology and entrepreneurship). He has chaired tracks at flagship SDN/NFV conferences, such as Carrier Network Virtualization, 5G Forum USA, NFV World Congress, and IoT World, among others. He manages the very active, 10,500+-member Carrier Ethernet LinkedIn Group. He earned his B.Tech. (EE) degree from IIT Kanpur, and his M.S. (signals and systems), M.S. (computer engineering), and Ph.D. degrees from the University of California, Santa Barbara.

DEEPAK KAKADIA is currently working in Mountain View, California, in the area of networking. Previously, from January 2013 to January 2015, he was team lead, Distinguished Member of Technical Staff, and IP network architect with Verizon Labs, leading network QoS analytics and network QoS optimization for LTE wireless network service provider networks in Palo Alto, California. From May 2005 to January 2013 he was with Verizon/Verizon Wireless in the headquarters of the Network Planning Group in Walnut Creek, California. Previously he was a staff engineer and IP network architect at Sun Microsystems Inc., Menlo Park, California, for a total of 11 years beginning in 1994. He also worked at Corona Networks as a principal engineer in the Network Management Systems group; Digital Equipment Corp, where he worked on DEC OSF/1; and Nortel Networks (Bell Northern Research) in Ottawa, Canada. He received a certificate in Networking from the Department of Electrical Engineering at Stanford University, Palo Alto, California. He has over 30 awarded patents and has filed over 20 additional patents in the areas of network and systems management and wireless technologies.

JENNIFER YATES is an assistant vice president at AT&T Labs-Research, leading the Networking and Service Quality Management (SQM) Research group, which is focused on inventing, prototyping, and driving new technologies for enabling operations transformation from a network to a service (or customer experience) focus, and enabling new services, enhancing reliability, and/or radically changing network costs. The group's innovative technologies are deployed across AT&T's networks and services. She also has an extensive publication record, has been a member of a number of TPCs (OFC, Sigcomm, NSDI, and others), is an Associate Editor for *IEEE/ACM Transactions on Networking*, and is the 2014/2015 Steering Committee Chair for *IEEE/ACM Transactions on Networking*. She received her Ph.D. from the University of Melbourne, Australia, and a B.E. (hons) and B.Sc. from the University of Western Australia. She was made an AT&T Fellow in 2013 for continued contributions in bridging network layers and management platforms to invent and deliver novel services and network management capabilities. She was honored with the AT&T Science & Technology Medal in 2006, and received the Victorian Photonics Network Achievement Award in 2004 and Top Young Innovators in Technology Review 100 in 2003.

AXEL CLAUBERG joined Deutsche Telekom AG in September 2011. Within the Group CTO team, he is responsible for DT's aggregation, transport and IP, and infrastructure cloud strategy. He has more than 30 years of experience in the IT and telecommunications industry. From 1998 until August 2011, he had various international leadership roles at Cisco Systems; his last role was sales CTO in Cisco's Emerging Markets theatre. Since December 2011, he has represented DT on the Open Networking Foundation Board of Directors. In 2014 he was appointed as an Advisory Director for the MEF Board.

MIRKO VOLTOLINI is VP of Technology, Architecture and Asset Management within Colt Network Services. He is responsible for technology and product development of Colt's network services, for the architecture of the Colt network, and for the management of Colt network assets. He joined Colt in 2002 and held several senior roles in the technical development and engineering area, covering data, voice, and IT services and technologies. Prior to joining Colt he worked at GTS/Ebone, Italtel, and ICT Consulting, an independent telecom consultancy company. He holds an M.Sc. in telecommunications engineering from Politecnico di Milano.

Orchestration of RAN and Transport Networks for 5G: An SDN Approach

Ahmad Rostami, Peter Öhlén, Kun Wang, Zere Ghebretensaé, Björn Skubic, Mateus Santos, and Allan Vidal

The authors present an overview of the benefits and technical requirements of resource coordination across radio and transport networks in the context of 5G. Then they discuss how software defined networking principles can bring programmability to both the transport and radio domains, which in turn enables the design of a hierarchical, modular, and programmable control and orchestration plane across the domains.

Abstract

The fifth generation of mobile networks is planned to be commercially available in a few years. The scope of 5G goes beyond introducing new radio interfaces, and will include new services like low-latency industrial applications, as well as new deployment models such as cooperative cells and densification through small cells. An efficient realization of these new features greatly benefit from tight coordination among radio and transport network resources, something that is missing in current networks. In this article, we first present an overview of the benefits and technical requirements of resource coordination across radio and transport networks in the context of 5G. Then, we discuss how SDN principles can bring programmability to both the transport and radio domains, which in turn enables the design of a hierarchical, modular, and programmable control and orchestration plane across the domains. Finally, we introduce two use cases of SDN-based transport and RAN orchestration, and present an experimental implementation of them in a testbed in our lab, which confirms the feasibility and benefits of the proposed orchestration.

Introduction

Similar to previous generation mobile communications systems, advances in technology and society are influencing how the next generation mobile network, the fifth generation (5G), is shaping up [1, 2]. With 3G and 4G, mobile traffic shifted from traditional telephony services to data, and building on this success, 5G aims to provide unlimited access to information by people and a large variety of connected devices. We will see a massive growth in both traffic and the number of connected devices. New services will be developed and launched in shorter time cycles than current networks allow. End-user services will continue to develop, but there will also be an increasing volume of machine-type communications with very different requirements on the network, from networks of sensors and actuators to performance-critical industrial applications. To ensure that networks will be able to cope with the diverse landscape of future services, a variety of forums like the Next Generation Mobile Network Forum (NGMN), International Telecommunication Union Radiocommunication Standardization Sector (ITU-R), and 5G Public Private Partnership (5G-PPP) have defined aggressive performance targets for 5G systems to fulfill future requirements, including access bit rates up to 10 Gb/s and a significant reduction in latency [2]. It should be noted, however, that the most demanding requirements will not apply to all services, but the network needs to be flexible enough to accommodate different services in a cost-effective manner.

The digital and mobile transformations currently sweeping through industries worldwide are giving rise to innovative cross-sector applications that are demanding in terms of network resources. Programmability and operational scalability are key enablers for rapid innovation, short time to market for deployment of services, and speedy adaptation to the changing requirements that modern industry demands. Furthermore, as end-to-end services are increasingly deployed in a distributed cloud environment, this programmability should span all relevant domains, including the radio access network (RAN), various network domains, and distributed processing, in an orchestrated manner.

Software-defined networking (SDN) [3] is a promising approach to bring the required programmability to different parts of the network; in particular, a lot of research has been done in introducing SDN to fixed networks covering different network applications. Recently, there have also been efforts to adopt these ideas in wireless networks [4, 5], although this area is less mature in comparison to fixed networks. Nevertheless, very little has been achieved when it comes to coordinated resource control across all interconnected domains, something that — as we elaborate on below — is a key aspect of 5G networks, and is the main focus of this article.

In the following sections we detail the relevance of the programmability in the context of 5G, and explain the foreseen benefit of coordination among the different domains of transport, RAN, distributed processing environment, as well as network and service functions. We then present a scalable orchestration architecture, and two different network scenarios experimentally showing benefits of cross-domain optimization. These example scenarios are based on a centralized RAN (C-RAN) deployment model where the radio equipment's functionality is split between remote radio units (RRUs) and baseband processing units (BBUs), which are interconnected using the high-speed common public radio interface (CPRI) through a fronthaul network.

Digital Object Identifier:
10.1109/MCOM.2017.1600119

Ahmad Rostami, Peter Öhlén, Zere Ghebretensaé, Björn Skubic, Mateus Santos, and Allan Vidal are with Ericsson Research; Kun Wang is with Acreo Swedish ICT and KTH Royal Institute of Technology.

Radio and Transport Interaction

Today, mobile networks are optimized for mobile broadband application. RAN and evolved packet core (EPC) functions are defined toward this backdrop even if additional applications are emerging, mostly in the Internet of Things (IoT) area (e.g., connected vehicles). Realization of these new use cases and services requires implementation of new features and capabilities in the network, which is a challenging and time-consuming task as it needs to go through lengthy standardization processes. This indicates that a more flexible and efficient way to add capabilities and customize deployments is needed, enabling network operators to support fast deployment of new services from a variety of applications and industries. It is already possible to share a network infrastructure among several mobile virtual network operators (MVNOs), each with its own business processes and customers. However, the MVNOs do not have the possibility to adapt to new features and capabilities in the network as required for new services.

In 5G, the concept of network slices is introduced, where each slice can span several segments of a network and be customized to support a specific service. Such services range from evolved mobile broadband and media distribution to different IoT applications, and even include applications and services that have yet to be defined. One important enabler for this is the decomposition of the mobile core functionality into granular functions and virtualization of them following the concept of network functions virtualization (NFV) [6, 7]. This enables flexible placement of the different virtualized network functions (VNFs) in centralized or distributed execution environments. For example, in a media distribution slice, core functions and caches could be placed close to a distribution location to optimize the performance. Efficient realization of this, however, requires coordination with the transport network, which provides connectivity among the VNFs.

Industrial remote control applications constitute another set of use cases, which have high requirements on the network in terms of availability, latency, and bandwidth. To fulfill end-to-end service requirements of these applications in a dedicated slice, each component of the slice needs to meet specific requirements. This includes processing performance, function placement, radio characteristics, and transport network.

Looking specifically into 5G radio, new challenges arise from a transport network perspective [8, 9]. In fact, already concepts are in development where the different domains of radio and transport can exchange performance data and optimize user experience by cross-domain optimization of traffic flows [10]. From the 5G deployment perspective, densification of the radio network by small cells implies that a user equipment (UE) will often be in the range of several radio base stations. Selecting a base station involves not only radio parameters, but also transport network performance. If a backhaul link experiences high packet loss, we would like to push services with higher requirements to different base stations with better transport connectivity. This type of load balancing requires information from both RAN and transport, and a UE may also be connected across different access technologies, which increases the need to coordinate between different technology domains.

In dense radio deployments, interference levels increase, which at times requires radio coordination capabilities for mitigation. However, the method used for handling interference depends on the deployment model. In a centralized baseband deployment, tight coordination features such as joint processing can be implemented at the cost of typically high CPRI bandwidths and stringent delay and jitter requirements. In traditional Ethernet or IP-based backhaul, tight coordination requires low-latency lateral connections between participating base stations.

In a traditional C-RAN architecture, the fronthaul connectivity is static. Introducing a flexible fronthaul network enables dynamic allocation of BBUs and RRUs, and a number of optimizations can then be applied [11], e.g.

- *Energy saving*: BBUs and RRUs can be put in a low-power state in times of low traffic.
- *Dynamic clustering*: To enable joint baseband processing, RRUs can be dynamically clustered into groups to optimize coordination gains.
- *Pooling*: Some scenarios allow for a reduced number of BBUs by flexibly allocating processing capacity to radio cells where demand is higher.
- *Shared fronthaul*: A fronthaul operator can share the network among several radio network operators to optimize use of the fiber infrastructure.
- *Resilience*: In cases of failure in BBU pools and/or transport connectivity, a coordinated mechanism is needed to restore the network to normal operation, or, if this is not possible, to at least ensure that basic coverage and operation are secured.

While the flexibility gain of such approaches has to be evaluated against the increased complexity, these different examples point to the value of increased flexibility and programmability, and more coordination among different parts of the network. Bringing the different domains together is a challenge from both the technical and operational perspectives. Later we adopt an SDN-based orchestration architecture spanning the different technology domains of transport, radio, and cloud to solve these issues. But before that, in the next section we present an overview of other requirements for developing a cross-domain orchestration architecture and enabling technologies.

The Need for Programmability

Meeting the objective of flexibility requires as a first step that the actual resources in the infrastructure can be adapted and changed dynamically without manual intervention by an operator. In addition, we need methods and procedures to develop applications and services on top of a flexible infrastructure, across different technology domains. To fulfill these requirements, a control architecture with a high level of *programmability* is required. Specifically, the control architecture should not be bound to a particular use case or scenario, and should enable a network operator to program customized algorithms into the control plane for optimization of RAN, transport, and cloud resources.

In addition to this overarching objective, the following features are required:

Meeting the objective of flexibility requires as a first step that the actual resources in the infrastructure can be adapted and changed dynamically without manual intervention by an operator. In addition, we need methods and procedures to develop applications and services on top of a flexible infrastructure, across different technology domains.

In a RAN, separation of data plane and control plane logic has been an important concept for a long time in the different technology generations. A somewhat more recent development is the self-organizing network framework, which uses a centralized SON controller to optimize network parameters and configurations on a coarse timescale.

Modularity: The control architecture should follow a modular architecture with well-defined control functions and interfaces. The interfaces and architectural building blocks should also support stacking in a recursive manner, to enable system deployments adjusted to specific scenarios.

Virtualization: The architecture should have the capability to divide physical and virtual resources of the infrastructure into separate groups (or slices) and allocate these to different clients. Here, clients can be higher-level controllers or service functions and applications. Dedicated slices should be isolated from each other for both security and performance reasons (e.g., to prevent an overloaded slice to negatively affect other slices).

Scalability: Control of resources in each of the domains is a complex task, as it usually requires dealing with a large number of network elements as well as that of control parameters and procedures. Therefore, joint control over the domains could easily become intractable, which should be avoided by proper design of the control architecture. Suitable abstraction methods are also needed to limit the complexity in higher layers and make the global optimization problems manageable.

To meet these requirements, we adopt SDN principles in the design of the overall control and orchestration architecture.

SDN

Programmability has been a hot topic in networking research for the last 25 years. Over the years several approaches, ranging from active networks to multiprotocol label switching (MPLS), have been explored for bringing more flexibility into the networks. SDN is the latest attempt in the quest for network programmability, and it has attracted much attention in both academia and industry. The SDN concept can bring the needed programmability in the transport part of mobile networks.

Through SDN, the main intelligence of the network control is decoupled from the data plane elements and placed into a logically centralized remote controller. This allows a network operator to directly program customized control algorithms into the network. A key concept of SDN is the abstraction of network elements (e.g., switches, routers, and access points) and specifying corresponding application programming interfaces (APIs) [3]. OpenFlow is an example of SDN abstractions, where switches' forwarding functionality is abstracted in the form of one or more flow tables, and the OpenFlow protocol specifies methods for programming the behavior of the tables by a remote controller [12].

SDN also enables creating multiple layers of abstractions on top of the controller in a recursive way. For example, a layer of abstraction on top of the SDN controller of a large network can hide all the details of that network and present the whole network as a big switch. This improves the modularity and scalability of the control architecture.

Programmable RAN

In a RAN, separation of data plane and control plane logic has been an important concept for a long time in the different technology generations. A somewhat more recent development is the self-organizing network (SON) framework, which uses a centralized SON controller to optimize network parameters and configurations on a coarse timescale (e.g., adjustment of power between neighboring radio base stations).

There is, however, a need for enhanced programmability in the RAN. One driver for this is the need for flexibility in supporting all different services in a timely manner. In research, there has been interest in SDN-like approaches in radio networks [4, 5]. In [4] a concept for software-defined fronthaul is presented along with some relevant use cases. It is also noted in [5] that mobile networks deal with a fundamentally different problem in complex radio environments, whereas SDN in packet networks mostly addresses forwarding. Consequently, solutions will be different, and [5] presents requirements for software-defined mobile networks.

In the transition to 5G, there are ongoing discussions on how to split the radio functionality. The current BBU/RRU and the associated CPRI may not be the best option going forward. Still, functions closer to the air interface with strict real-time characteristics should continue to be deployed on specialized hardware. But other functions could be deployed on general-purpose hardware, possibly in a virtualized environment [13]. Add to this the possible placement of mobile core and service-specific functions on distributed execution environments, possibly co-located with radio functions. Then the need for coordination between dedicated radio hardware, network functions, processing environment, and transport connectivity becomes more obvious.

Cross-Domain Orchestration Architecture

In this section, we present a hierarchical cross-domain orchestration architecture, which follows the SDN principles and addresses the above-mentioned challenges for programmability and flexibility. The proposed architecture is depicted in Fig. 1. At the bottom level of the architecture we have heterogeneous sets of resources distributed across different domains. These include radio, transport, as well as cloud (compute and storage) resources. The radio resources are primarily at the access edge of the network. The transport resources are distributed across different parts of the network (e.g., access and aggregation) and usually encompass different technology domains like packet, optical, and microwave networks. The cloud resources are also distributed across the access/aggregation and core of the network for hosting various VNFs.

Resources within individual domains are controlled through a domain-specific controller in a programmable way. A domain controller — through a programmatic northbound API — exposes the domain capabilities to higher-layer controllers/orchestrators, and enables them to dynamically program the corresponding resources. Typically, a controller exposes an abstract view of the domain resources over the API, and hides most of the details. Designing the abstraction layer on top of a domain typically involves a trade-off between the optimal resource utilization and the simplicity and scalability of the operation and control. Specifically, exposing more detailed information of domain resources enables higher layers to make optimized resource allocations, but on the other hand increases the complexity of higher-layer controllers/orchestrators as they will need to deal with lots of information and updates

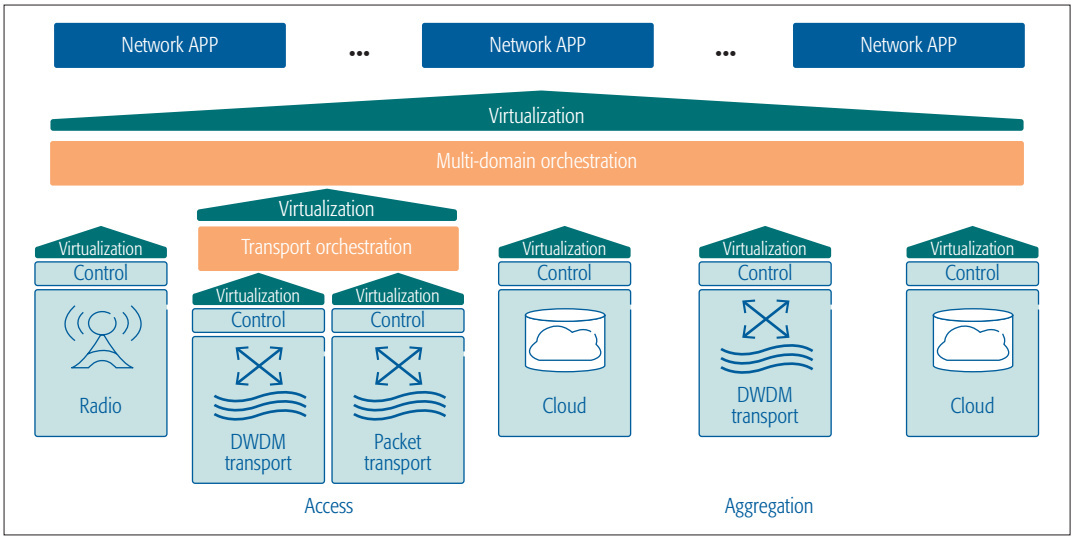


Figure 1. Hierarchical cross-domain orchestration architecture.

[14]. Additionally, a controller can take its client's needs as input for deciding the level of abstraction. Obviously, the level of abstraction requested by a client should be within the policy defined by the operator. Furthermore, the controller might virtualize the resources and allocate slices with different levels of abstractions to higher-layer clients.

To elaborate on the domain-specific controller, let us make an example. Consider a wavelength-switched optical transport network, where a centralized SDN controller manages the network resources as described in [11]. The SDN controller has comprehensive knowledge of the network topology and optical resources (e.g., available wavelengths and transceivers), and implements a routing and wavelength assignment algorithm. On its southbound interface the controller communicates with the optical devices to configure the wavelength switching tables. And on its northbound API the controller hides all optical details and presents the whole domain as a single optical cross-connect (OXC) whose ports are the ingress/egress ports of the optical domain.

On top of the domain controllers, there are one or more layers of orchestration — usually separated into service and resource orchestration layers. The service orchestrator takes high-level definitions of services and applications, and translates these into lower-level components using a catalog with pre-defined building blocks. It also handles life cycle management with tasks like service deployment and upgrades. It interfaces to one or several resource orchestrators, where the actual resources are handled for realizing the service. The resource orchestration layer combines resources of the same or heterogeneous types across multiple domains into a unified resource representation. For example, a transport orchestration layer in the access segment of the network combines abstract views of resources from multiple transport domains (e.g., packet and optical networks) to create a unified and technology-agnostic presentation of transport resources in that segment (Fig. 1). Similarly, at the topmost layer of the resource orchestration radio, the transport and cloud domains are combined to create a unified abstraction of resources. The unified abstraction is then exposed directly toward network applications or the service orchestration over a programmatic API. In the following, we focus only on the resource orchestration functions and how network applications can dynamically program the resources, as needed, through this API.

Also note that as we move up the control/orchestration layers, the resource presentations become more abstract. Therefore, a key function of an orchestration layer is mapping the requests coming from a network application or a higher-layer orchestrator to lower-layer orchestration and control. In other words, requests arriving at the topmost layer of orchestration are expressed at a very abstract level, and they are translated into more concrete configuration requests as they are passed down the control plane layers, until the lowest-layer controllers translate them to domain-specific configuration commands. The mapping of service requests to the resources should on one hand fulfill the performance requirements of the services (e.g., in terms of bandwidth, latency, and compute resources), and on the other hand target the optimization of the resource utilization across the network.

Additionally, a virtualization layer on top of a resource orchestrator enables the unified set of resources to be sliced and allocated to different network applications. As discussed before, this enables efficient sharing of an operator's resources among several network applications or service providers.

A key aspect of the orchestration architecture is to specify the abstractions and interfaces between different control and orchestration layers. While there are many approaches for abstracting resources of individual domains, designing combined abstraction models for heterogeneous resources is a challenge, and is a subject of ongoing research. For example, the European project UNIFY has proposed a joint abstraction of networking and compute resources in the form of a big-switch big-software (BiS-BiS), which presents a virtualization of a networking element connected with a compute node [15]. In this model, the resource requests between the resource orchestration layers can be recursively expressed as network function forwarding graphs (NF-FGs), where an NF-FG presents a mapping of a group of NFs

The mapping of service requests to the resources should on one hand fulfill the performance requirements of the services, e.g. in terms of bandwidth, latency and compute resources, and on the other hand target the optimization of the resource utilization across the network.

The programmability feature is realized through the orchestration's northbound API, which allows network applications to implement customized optimizations across radio, transport and cloud domains. As for the modularity, adding a new technology domain to the control plane is straightforward, and does not require changes to the existing parts.

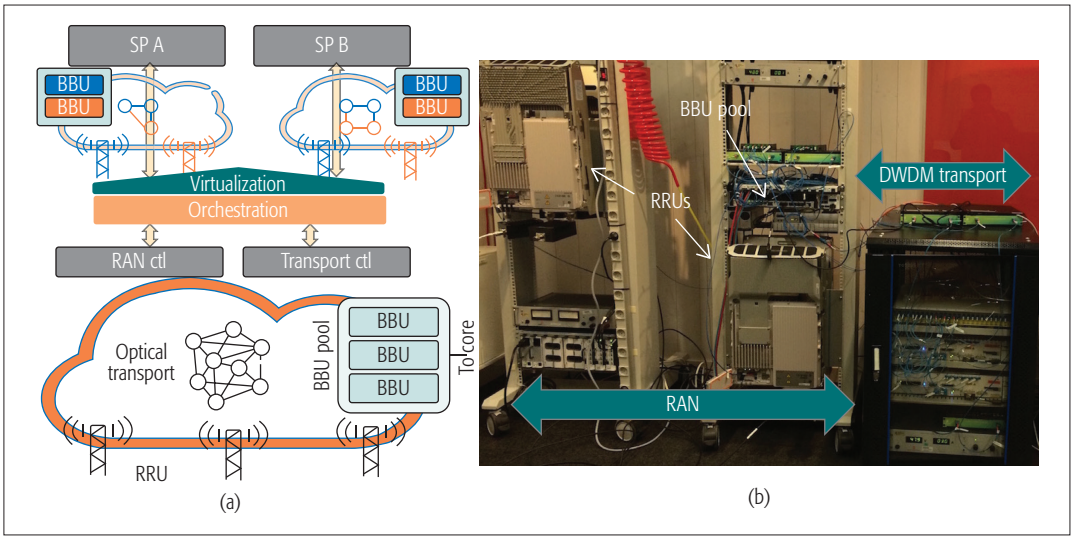


Figure 2. a) Architecture of the RAN-transport orchestration testbed; b) a snapshot of the experimental testbed showing RAN and optical transport domains.

and their corresponding forwarding overlay into the abstract view of the infrastructure (i.e., BiS-BiS) [15]. In the next section, we explain how we adopt this concept for creating joint abstractions of radio and transport resources.

Let us now elaborate on how the presented control architecture meets the requirements mentioned in the previous section. The programmability feature is realized through the orchestration's northbound API, which allows network applications to implement customized optimizations across the radio, transport, and cloud domains. As for modularity, adding a new technology domain to the control plane is straightforward and does not require changes to the existing parts. Also, a service provider can perform the cross-domain orchestration and optimization of resources on top of any combination of owned and leased domains, thanks to the virtualization on top of every control/orchestration layer. Finally, scalability is supported through several layers of abstractions, where specific details of each domain are hidden below corresponding abstractions.

Use Cases and Proof of Concept

In this section we present two use cases of the cross-domain orchestration, which we have experimentally demonstrated in a testbed in our lab. The objective is to demonstrate the feasibility as well as benefits of the designed architecture. The use cases particularly focus on the RAN-transport orchestration to fulfill the requirements of 5G. However, for the sake of simplicity, in our testbed we utilize existing 4G radio access points and a mobile core network. The first use case presents sharing of joint RAN-transport resources between two service providers (SPs), and the second one demonstrates how an SP can customize its own slice.

Radio-Transport Orchestration Testbed

Our testbed is composed of the following two domains:

- Mobile broadband domain: provides broadband services to mobile users employing LTE technology. The domain is composed of a group of LTE access points, deployed according to the C-RAN architecture. The mobile network relies on wavelength connectivity services of the transport domain for CPRI transport.
- Optical transport domain: is a dynamic wavelength routed network based on dense wavelength-division multiplexing (DWDM) and provides programmable fronthaul services to the mobile network at the wavelength level. The domain is composed of optical DWDM switches, optical add/drop multiplexers, and tunable transceivers.

Figure 2 depicts the architecture of the testbed together with a snapshot of the physical equipment deployed for this purpose. The control/orchestration plane of the testbed is realized following the architecture presented earlier.

The transport domain is controlled by the open source SDN controller OpenDaylight [16]. The controller has been customized to support circuit-switched network control and optical path computation, and a southbound plugin has been developed to allow configuration of the optical switches over their existing command line interfaces (CLIs). Also, an abstraction layer on top of the controller is implemented, which adopts the BiS-BiS model and presents the transport domain as a large OXC over an NF-FG interface toward the orchestrator (Fig. 3). For the mobile domain we use an existing CLI-based RAN manager that centrally controls the RAN resources. The RAN control functions include activation and configuration of cells, assignment of BBU resources to RRUs, as well as management of users' handovers among cells. We model the functionality of RRUs and each baseband processor (i.e., BBU) as individual NFs, which enables us to abstract the RAN resources as a BiS-BiS model. Then the orchestrator combines the BiS-BiS models of the RAN and transport domains into a unified model. The orchestrator with the NF-FG interfaces has been realized in Python.

Scenarios and Results

The first use case demonstrates sharing of a mobile access network infrastructure between two service providers. For realizing the infra-

structure sharing, the orchestrator's virtualization layer implements joint slicing of resources in the RAN and transport domains. The radio resources included in the slicing are RRUs and BBUs. Figure 2a illustrates an example of the overall view of the resources at the infrastructure provider level, and the virtualized view presented to any of the two SPs (the smaller clouds in the figure). There are two types of virtualized resources: dedicated (shown in blue in Fig. 2a) and shared (orange). While dedicated resources are guaranteed to be at the disposal of an SP at all service operation times, the shared resources can be used by either of the SPs at any time. The orchestrator ensures the isolation between the dedicated resources of each SP and also resolves possible conflicts for using the shared resources according to a sharing policy. The slicing of the resources with a first-come first-served sharing policy is successfully realized in our testbed.

In the second use case, we demonstrate how SP A can run a customized optimization of radio and transport resources within its allocated slice. The infrastructure provider's orchestration is configured to provide a group of dedicated transport and radio resources to SP A as illustrated in Fig. 3. These include:

- Four RRUs distributed across a residential and a business area. Each area is equipped with two RRUs: a macrocell and a small cell. While the macrocell provides the coverage across an area, the small cell is used for providing additional capacity in the area if needed
- A pool of three BBUs
- Three optical connections abstracted as a 3:4 OXC, which can be dynamically reconfigured to connect any of the three BBUs to any of the idle RRUs

SP A utilizes the allocated resources to create an elastic mobile broadband (EMBB) service, where the service capacity is dynamically and automatically scaled up and down — when and where needed. The trigger for scaling the service capacity comes from live monitoring of the service demand in the RAN. The monitoring is performed through measuring the throughput of active cells by the RAN controller, upon request from the SP. The EMBB service logic has three states (Fig. 4a). In state 1 (the default state), only the two macrocells are activated, providing coverage in both areas. When extra demand is identified by the EMBB service manager (residing inside SP A), the corresponding small cell is activated (state 2 or 3). To do that, the service manager requests the orchestrator to reprogram the testbed to adjust the service capacity through de/activating small cells. The orchestrator translates the request into required configurations in the transport and RAN, and the configurations are applied in the data plane by the corresponding controllers. For example, activation of a new RRU involves assigning and configuring appropriate BBU resources from the BBU pool, and establishing the wavelength connectivity between the RRU and the assigned BBU. Note that only one small cell is activated at a time to serve the area with a higher demand. This demonstrates dynamic reuse of resources where a four-cell RAN requires only three optical connections and three BBU resources. In a real deployment, there would be many more RRUs and BBUs. Our numerical analysis, although not presented here, indicates that implementing the EMBB service in a metro area with hundreds of RRUs leads to a saving of around 30 percent in terms of required transport and radio resources [17].

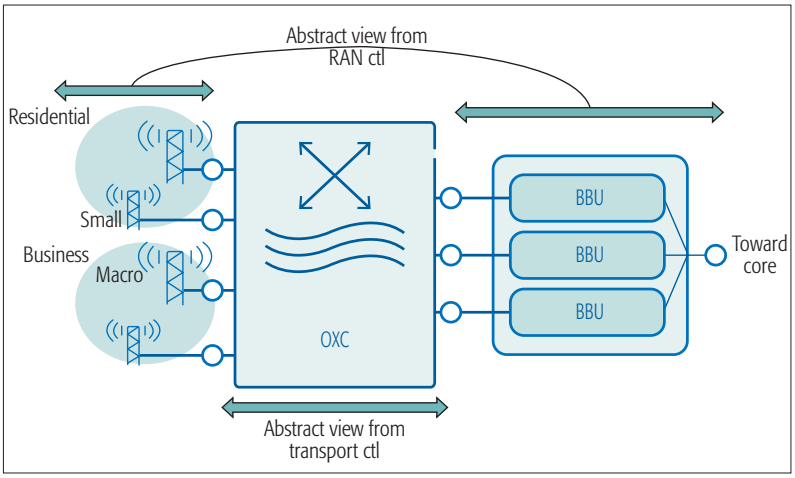


Figure 3. Abstract view of resources exposed by the infrastructure provider's orchestrator toward SP A. The orchestrator creates the abstract view by combining views received from transport and RAN controllers.

Figure 4b shows a representative example of traffic measurements performed during the runtime of the EMBB service in our testbed. The presented traffic measurements clearly show the transition among states 1, 2, and 3 over the runtime of the system.

The use cases above demonstrate the value of a global cross-domain orchestration for both an infrastructure provider and its customers (i.e., SPs). The infrastructure can share its resources among different customers, leading to better resource utilization and higher revenues. At the same time, the customers can run their own customized control mechanisms, leading to a much shorter time to scale services or to create new ones.

Conclusion

Fast introduction of new services and dynamic scaling of them are among major expectations on future networks in general and 5G in particular. Flexibility in all networking domains is a crucial requirement to fulfill these expectations. Furthermore, coordinated control of heterogeneous resources across multiple domains is needed, and SDN is a promising approach for bringing the flexibility and realizing the required coordination. In view of this, we present an SDN-based cross-domain orchestration architecture and validate it by experimental realization of two use cases. The use cases, which are based on joint slicing of RAN-transport resources and the EMBB service, demonstrate the benefits of the joint orchestration for both infrastructure providers and service providers. Specifically, service agility and efficient resource utilization are among the main benefits of the proposed orchestration architecture.

Acknowledgment

This work was supported in part by the VINNOVA project "Kista 5G Transport Lab (K5)."

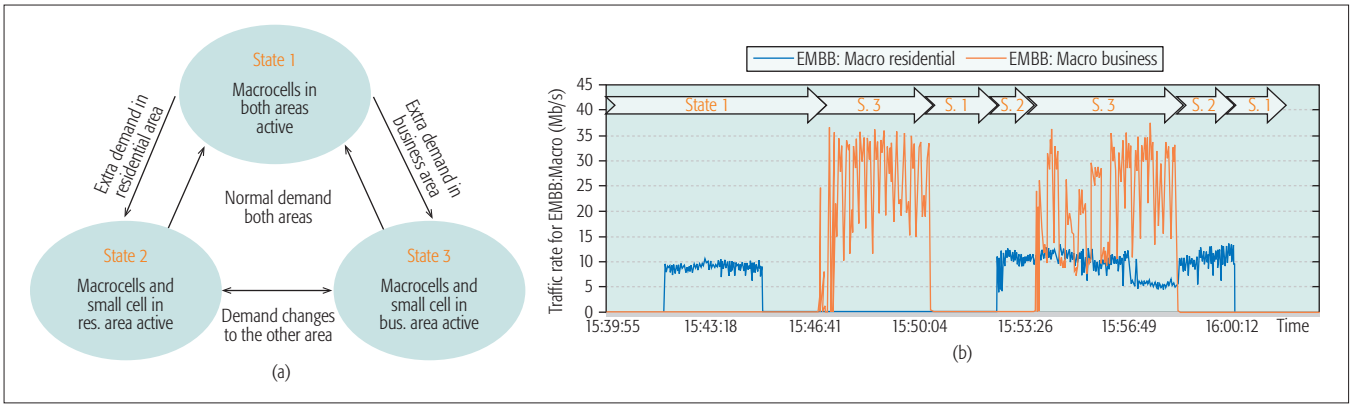


Figure 4. a) State diagram of the EMBB service; b) traffic measurements of the service: the demand threshold to activate a small cell is set to 10 Mb/s. The arrows indicate states (transition) during the service operation. UEs connected to the macro RRUs are configured to download large videos to trigger state transition.

References

[1] J. F. Monserrat *et al.*, "Rethinking the Mobile and Wireless Network Architecture: The METIS Research into 5G," *Proc. Euro. Conf. Networks and Commun.*, June 2014, pp. 1–5.

[2] E. Dahlman *et al.*, "5G Radio Access," *Ericsson Technology Review*, June 2014, pp. 1–7.

[3] D. Kreutz *et al.*, "Software-Defined Networking: A Comprehensive Survey," *Proc. IEEE*, vol. 103, no. 1, Dec. 2014, pp. 14–76.

[4] M. Y. Arslan, K. Sundaresan, and S, Rangarajan, "Software-Defined Networking in Cellular Radio Access Networks: Potential and Challenges," *IEEE Commun. Mag.*, vol. 53, no. 1, Jan 2015, pp. 150–56.

[5] T. Chen *et al.*, "Software Defined Mobile Networks: Concept, Survey, and Research Directions," *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015, pp. 126–33.

[6] J. Kempf *et al.*, "Moving the Mobile Evolved Packet Core to the Cloud," *Proc. Wireless and Mobile Computing, Networking and Commun.*, Oct. 2012, pp. 784–91.

[7] P. Gurusanthosh, A. Rostami, and R. Manivasakan. "SDMA: A Semi-Distributed Mobility Anchoring in LTE Networks," *Proc. Mobile and Wireless Networking*, Aug. 2013, pp. 133–39.

[8] P. Öhlén *et al.*, "Data Plane and Control Architectures for 5G Transport Networks," *J. Light. Tech.*, vol. 34, no. 6, Mar. 2016, pp. 1501–08.

[9] B Skubic *et al.*, "Rethinking Optical Transport to Pave the way for 5G and the Networked Society," *J. Light. Tech.*, vol. 33, no. 5, Mar. 2015, pp. 1084–91.

[10] S. Dahlfort *et al.*, "Radio Access and Transport Network Interaction — A Concept for Improving QoE and Resource Utilization," *Ericsson Technology Review*, July 2015, pp. 1–8.

[11] A. Rostami *et al.*, "First Experimental Demonstration of Orchestration of Optical Transport, RAN and Cloud Based on SDN," *Proc. OFC Post-Deadline Papers*, Mar. 2015, paper TH5A.7.

[12] N. McKeown *et al.*, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 38, no. 2, Apr. 2008, pp. 69–74.

[13] E. Westerberg, "4G/5G RAN Architecture: How a Split Can Make the Difference," *Ericsson Tech. Review*, vol. 93, no. 6, July 2016.

[14] M. Fiorani *et al.*, "Transport Abstraction Models for an SDN-Controlled Centralized RAN," *IEEE Commun. Lett.*, vol. 19, no. 8, Aug. 2015, pp. 1406–09.

[15] EU Project Unifying Cloud and Carrier Networks; https://www.fp7-unify.eu/, accessed July 2016.

[16] Linux Foundation OpenDaylight Project; https://www.opendaylight.org/, accessed July. 2016.

[17] M. R. Raza *et al.*, "Demonstration of Dynamic Resource Sharing Benefits in an Optical C-RAN," *IEEE J. Opt. Commun. Net.*, vol. 8, no. 9., Sept. 2016.

Biographies

Ahmad Rostami (ahmad.rostami@ericsson.com) is a senior researcher in networking technologies at Ericsson Research, where he leads activities in the area of programmable networks as well as control and orchestration architectures and protocols for 5G networks. Before joining Ericsson in 2014, he worked at the Technical University of Berlin (TUB) as a senior researcher and lecturer. At the university his areas of research covered network control and SDN technologies. He holds a Ph.D. (summa cum laude) in communication networks from TUB, and an M.Sc. in electrical engineering (communication networks) from Tehran Polytechnic.

Peter Öhlén is a principal researcher at Ericsson Research. He received an M.Sc. in engineering physics from the Royal Institute of Technology (KTH), Sweden, in 1995. In 2000 he received a Ph.D. in photonics, also from KTH. He has been with Ericsson since 2005. With more than 15 years of experience in telecommunications, he has worked with research and development in transport networks, network control, SDN, fiber access technologies, fiber optic transmission, radio networks, optical and electronic subsystem design, simulation methods, and project and program management. He was heavily involved in the standardization of 10Gb Ethernet and in the FSAN group for standardization of XG-PON systems. His current research focuses on network control, cross-domain orchestration, and 5G transport networks.

Kun Wang received his M.Sc. in electrical engineering from KTH in 2007. He has been an active member of European research projects like OASE, Alpha, and MUSE. Currently, he is working at Acreo Swedish ICT and KTH toward his Ph.D. in next generation optical transport networks. His research interests include FTTx networks, software defined networking, C-RAN, and 5G mobile network.

Zere Ghebretensaé graduated from the Institute of Technology Linköping, Sweden, with an M.Sc. degree in technical physics and electronics. Since then he has worked at Televerket Radio and Telia Research with radio and optical fiber transmission research. He joined Ericsson Research in 2000, and started working in packet and optical networking and later as project leader for various small cell backhaul research activities. He has participated and monitored OIF, ITU, and IEEE work, and was work package leader of the FP6 MUSE project and a task leader of the demonstration part of the FP7 COMBO project.

Björn Skubic is a senior researcher in networking technologies at Ericsson Research. He joined Ericsson in 2008, and has worked in areas such as optical transport, energy efficiency, and fixed access. He has a Ph.D. in physics from Uppsala University, an M.Sc. in engineering physics from KTH, and an M.Sc. in business administration and economy from Stockholm University.

Mateus Augusto Silva Santos received his M.Sc. (2009) in computer science and a Ph.D. (2014) in electrical engineering from Universidade de São Paulo, Brazil. From 2013 to 2014 he was a research scholar with the Inter-Networking Research Group at the University of California Santa Cruz. He was also a postdoctoral researcher with the University of Campinas. His research interests are in software-defined networking, network security, and wireless networks. He has industry experience at Hewlett-Packard and EMBRAER. He is currently a researcher at Ericsson in Brazil.

Allan Vidal received his M.Sc. in computer science from Universidade Federal de São Carlos in 2015. He has previous experience as a researcher and developer at CPqD and Lenovo, and is currently a researcher at Ericsson in Brazil. He has worked on open source SDN projects such as RouteFlow and libfluid. His research interests are in software-defined networking, network management, and data plane programmability.

SDN-Based IP and Layer 2 Services with an Open Networking Operating System in the GÉANT Service Provider Network

Pier Luigi Ventre, Stefano Salsano, Matteo Gerola, Elio Salvadori, Mian Usman, Sebastiano Buscaglione, Luca Prete, Jonathan Hart, and William Snow

Abstract

The migration of service providers' wide area networks toward SDN is a challenging task. In this article, we consider the critical requirements of GÉANT, the 500 Gb/s pan-European provider interconnecting 38 national research and educational networks, for a total of 50 million users. A long-term evolution path toward the *softwarization* of GÉANT is discussed, consisting of four steps to be realized in future years, from providing SDN-based connectivity, to the so-called software defined infrastructure (SDI). As a first step, the softwarization of some basic services currently offered by GÉANT is considered: *GÉANT IP*, *GÉANT plus*, and *GÉANT Open*. This article reports the concrete experience in the SDN-based design and implementation of these services, which have been called L3-SDX and L2-SDX. Both use cases have been addressed with the use of the open source Open Network Operating System (ONOS®). The L3-SDX service has been developed on top of an existing ONOS application, called SDN-IP. SDN-IP allows interconnections between SDN and legacy networks through BGP. The L2-SDX service has been realized as a new ONOS application. Both services are currently deployed on the GÉANT Testbed Service, a continental facility offering geographical virtual testbeds to the research community. The article reports the experience gained from this experimental deployment and discusses the benefits for a service provider like GÉANT.

Introduction

Software defined networking (SDN) is a recent paradigm [1] potentially able to transform the design of both data center and wide area networks. The promise of SDN is to foster innovation and flexibility thanks to centralized network control and standard interfaces. The fundamentals of the SDN approach are:

- The separation of control and data planes
- The logical centralization of the former as a software layer called controller or network operating system (NOS)
- The introduction of a flexible forwarding paradigm (based on filtering matches and actions)
- The direct control of the hardware through common management interfaces (e.g., OpenFlow)

SDN can be seen as part of an even wider trend toward the softwarization of networks [2, 3], which implies a complete rethinking of how service provider networks are now structured. It is expected that this process will greatly increase the flexibility and efficiency of networks, reducing equipment and operational costs.

In this article, we start from the analysis of current services offered by GÉANT, the 500 Gb/s pan-European network interconnecting 38 national research and educational networks (NRENs), for a total of 50 million users. We refer to the NRENs as the *customers* of GÉANT. We identify the GÉANT needs and requirements toward the upgrading of its infrastructure. A long-term evolution path toward the softwarization of GÉANT is discussed, consisting of several steps to be realized in future years, from providing SDN-based connectivity services to the so-called software defined infrastructure (SDI) [4, 5], which is also able to dynamically offer a wide range of computing/storage/network resources.

The first step of the GÉANT migration process consists of the *SDNization* of some operational services. In particular, we consider *GÉANT IP*, which is the basic service providing Internet connectivity to the NRENs, and two layer 2 connectivity services called GÉANT plus and GÉANT open. These services are currently delivered through 26 points of presence (PoPs) located in Europe and 2 open exchange points (OXPs) in London and Paris (see the next section for further details). The OXPs are similar to the standard Internet exchange points (IXPs), allowing NRENs to exchange traffic with external (non-GÉANT) networks. The introduction of SDN technologies in an IXP is referred to as software defined (Internet) exchange (SDX) [6]. In this article, we give a wider meaning to the SDX concept, extending its potential applicability not only to the exchange points (IXPs or OXPs) but also to the PoPs. We designed and developed two SDN-based services called L3-SDX and L2-SDX (L3 and L2 stand for layer 3 and layer 2). These services represent the fundamental building blocks of the augmented

The migration of service providers' wide area networks toward SDN is a challenging task. The authors consider the critical requirements of GÉANT, the 500 Gb/s pan-European provider interconnecting 38 national research and educational networks, for a total of 50 million users. A long-term evolution path toward the softwarization of GÉANT is discussed, consisting of four steps to be realized in future years.

Pier Luigi Ventre is with GARR/University of Rome Tor Vergata; Stefano Salsano is with University of Rome Tor Vergata/CNIT; Matteo Gerola and Elio Salvadori are with Create Net; Mian Usman and Sebastiano Buscaglione are with GÉANT; Luca Prete, Jonathan Hart, and William Snow are with ON.LAB.

Digital Object Identifier: 10.1109/MCOM.2017.1600194

The GÉANT infrastructure offers extensive links to networks in other world regions. External peers (other NRENs and external Autonomous Systems) are interconnected through 26 POPs, located all over Europe, and two Open eXchange Points (OXPs). GÉANT offers a wide range of connectivity and network management services

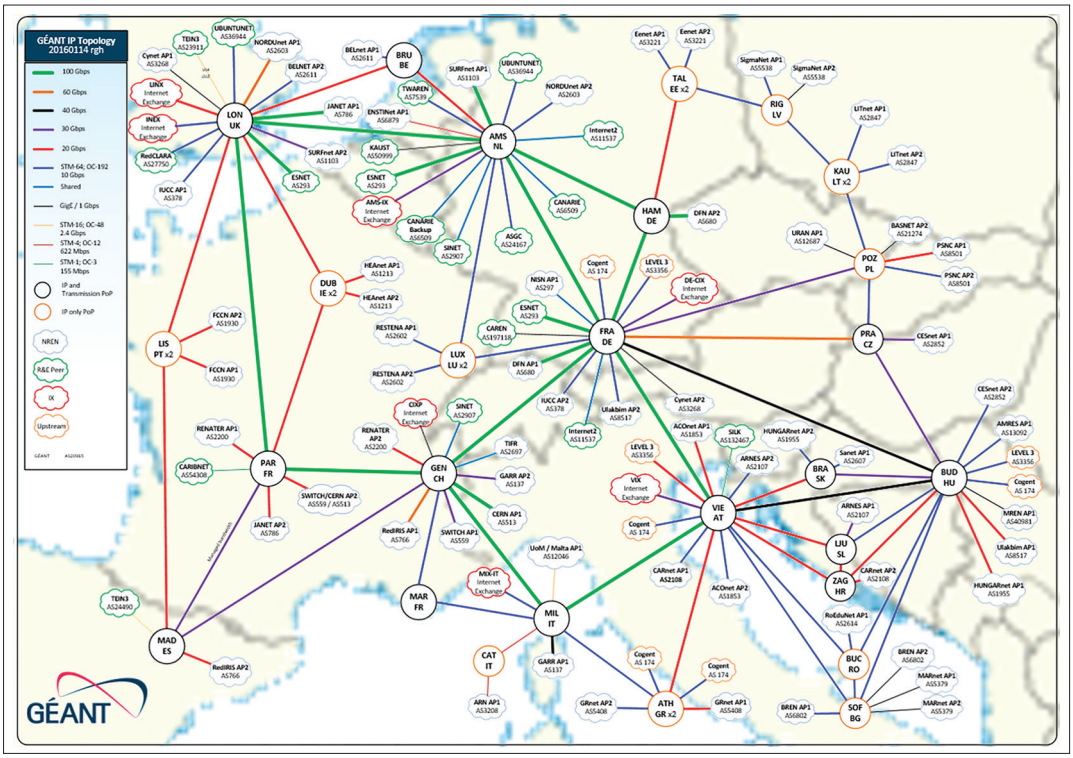


Figure 1. GÉANT IP topology and GÉANT OXPs.

SDX mentioned above and have been developed using Open Network Operating System (ONOS) [7], a promising open source solution for the SDN control plane.

GÉANT Network and Services

As one of the largest and most complex research and education networks in the world, GÉANT needs to support different services, such as standard IP transit connectivity and ultra-high-capacity data center interconnections. The GÉANT infrastructure offers extensive links to networks in other world regions. External peers (other NRENs and external autonomous systems) are interconnected through 26 POPs, located all over Europe, and two OXPs (Fig. 1). GÉANT offers a wide range of connectivity and network management services as described in [8]. We focus on a subset of them, called GÉANT IP, GÉANT plus, and GÉANT Open.

GÉANT IP provides IP transit services to interconnect participating NRENs together and with other approved research organizations and providers. It provides a peering service for IP traffic, isolated from general-use Internet access. GÉANT plus offers the NRENs point-to-point L2 (Ethernet) circuits among endpoints at GEANT PoPs [8]. The PoPs constitute the backbone of the dual (optical transmission and packet) layer network through which GÉANT supplies connectivity to its customers. From the GÉANT perspective, the PoPs are the endpoints of GÉANT IP and GÉANT plus. Finally, with the GÉANT Open service, NRENs can connect with external (non-GÉANT) networks through the OXPs. Inside an OXP, the customers (NRENs or external participants) request the establishment of L2 circuits between endpoints, which are manually provisioned through virtual LAN (VLAN) tunnels. The customers can use the L2 circuits for whatever reason, including private Border Gateway Protocol (BGP) peering. Therefore, OXPs are different from traditional IXPs, which provide a switched L2 infrastructure used by multiple participants to exchange traffic through public BGP peering.

Toward a New Service Development and Provisioning Approach

The current *connectivity and network management services* of GÉANT are mostly based on traditional IP/MPLS (multiprotocol label switching) control plane architectures running on top of complex and expensive proprietary equipment. In most cases, the services rely on proprietary software and specific vendor solutions, making it hard to innovate and offer new services. The management of a large-scale network like GÉANT is largely based on proprietary (and expensive) tools, which again constitute a barrier to innovation. A second issue is that the service provisioning phase often includes manual operations, resulting in provisioning times on the order of days. In such a scenario, the introduction of SDN and in general of *softwarization* can bring substantial benefits:

- Provisioning procedures can be drastically simplified.
- Cheaper hardware could replace current equipment.
- The openness of the SDN approach avoids the need for complex distributed control plane architectures and reduces the number of running protocols.
- Proprietary implementations can also be avoided or reduced, mitigating interoperability problems and migration issues.

Softwarization facilitates the development and introduction of new services of strong interest

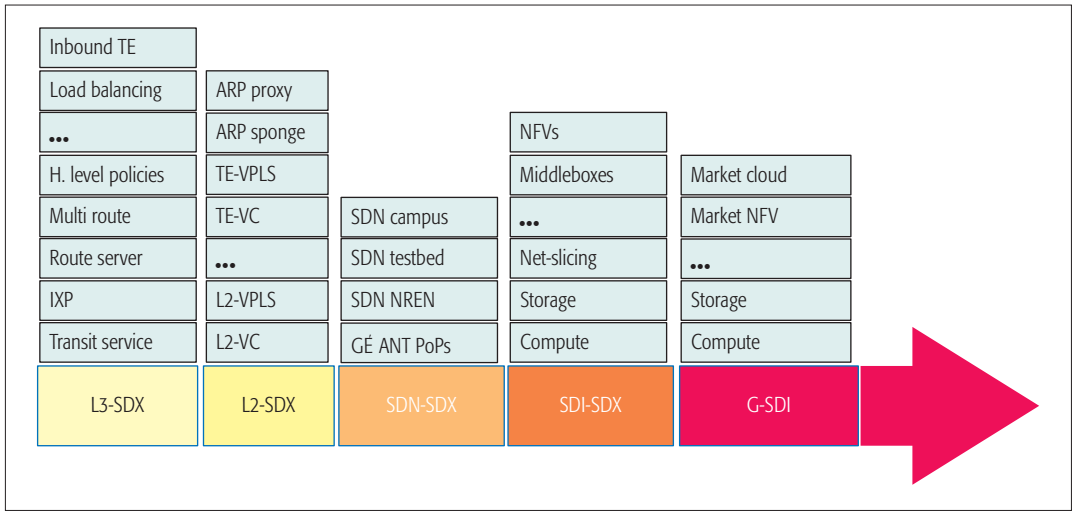


Figure 2. Softwarization path.

The migration of the GÉANT network and its services toward an SDI is a challenging task, which cannot happen overnight. We decomposed the transition of GÉANT to the SDN paradigm in incremental steps. Each step enhances the GÉANT infrastructure, making it more sustainable, more manageable, and less expensive, and introduces new services/functionalities to the portfolio.

to GÉANT but difficult to implement using the current technologies, such as application- specific peering, inbound traffic engineering, load balancing, and steering of traffic for service function chaining [9].

GÉANT Softwarization Path

The migration of the GÉANT network and its services toward an SDI is a challenging task, which cannot happen overnight. We decomposed the transition of GÉANT to the SDN paradigm in incremental steps, as shown in Fig. 2. Each step enhances the GÉANT infrastructure, making it more sustainable, more manageable, and less expensive, and introduces new services/functionalities to the portfolio. The transition path also takes into account the operational requirements of a production network. This migration strategy rationalizes and extends some ideas already presented by Monga in [5]. The idea behind the strategy is to initially introduce the concept of SDX, replacing the current architecture based on PoPs and OXPs, and then to progressively enhance its functionality. The first step of the path is the realization of the L3–SDX and L2–SDX, referred to as *pure connectivity* SDX [5, 10]. L3-SDX supports GÉANT IP, while L2-SDX represents the SDNization of GÉANT plus and GÉANT Open. With this first step, there is no full migration to SDN technology, because NRENs can run legacy technologies without direct interaction with the GÉANT SDX operations.

Assuming that the NRENs will set up their own SDN infrastructures, the next step, called SDN–SDX, consists of the interconnection and harmonization of the SDN infrastructures between GÉANT and the NRENs. To understand the advantages of this step, consider that the NRENs have their customers (research organizations), which requires connectivity toward other research organizations in the same NRENs, in remote NRENs, or outside the GÉANT's NRENs. Thanks to the SDN-SDX, the NRENs and the research organizations could fully exploit the advantages of the SDN paradigm, leveraging end-to-end SDN-based services spanning the GÉANT backbone and the NRENs.

The next step, denoted as SDI–SDX, refers to a GÉANT SDN infrastructure augmented with cloud resources: an NREN can request not only networking services, but also compute and storage resources, outsourcing some of its computations to the SDX cloud.

The final step is referred to as G-SDI, for global SDI. It foresees a wider adoption of the full softwarization (SDN augmented with storage and computing resources) by all NRENs. Following this approach, an end user (research organization) can obtain compute and storage resources from other NRENs or from GÉANT, leveraging the resources offered through a logically global GÉANT SDX. At this step, considering GÉANT's position in the European scenario, it will be possible to exploit the Buyya vision of market-oriented cloud computing [11]. GÉANT's role fits well as a "super party" that manages the market. This evolution of the GÉANT infrastructure encompasses an economic model, useful for the auto-sustainability of the GÉANT project and its participants (the NRENs themselves).

GÉANT Requirements for SDN-Based IP and Layer 2 Services

L3-SDX and L2-SDX have been identified as the first step of the GÉANT softwarization path. We carried out a thorough analysis of the requirements on these services from the perspective of the GÉANT service provider, summarized in Table 1. The requirements are classified as functional, non-functional (i.e., referring to performance or reliability), and operational (e.g., related to monitoring or logging).

Based on this requirement analysis, we selected ONOS [7] as the controller platform. In particular, the existing SDN-IP ONOS application provided a very good fit with the functional requirement of L3-SDX, and the ONOS resilience and distribution features provided a good match with the identified non-functional requirements.

ONOS: SDN Network Operating System for Service Providers

ONOS is an open source SDN control plane platform, meeting service provider requirements, released in 2014 as an open source project by

ON.Lab. ONOS provides a stable implementation of a scalable, highly available, and resilient network operating system (NOS).

The overall system has been conceived as a distributed system in the form of a cluster, composed of multiple instances, all functionally identical to each other. The architecture (Fig. 3) can be structured in three tiers: a protocol-aware southbound (SB) layer, a protocol-agnostic distributed core layer, and an application layer. Each tier is a collection of pluggable modules/subsystems realizing specific functionalities that make up the ONOS platform. An application programming interface (API) is exposed at each tier, providing isolation and modularity.

The distributed core is responsible for synchronization and coordination between the instances in the cluster. It builds a global network view based on information learned on the SB API and offers services to the application layer. In order to achieve scalability and provide resiliency, various distribution mechanisms are available through a set of primitives. Each core subsystem uses these primitives in different ways according to the consistency requirements of the state it is managing. On top of the distributed state, a logically centralized network view is constructed and presented to applications. In addition, work is partitioned among the instances in the cluster. For example, each instance is elected to be responsible for managing a subset of the devices in the network, while the other instances are ready to step in if the primary instance fails. In the case of data plane failures, built-in mechanisms for traffic rerouting are activated.

Requirement	Service	Type	Priority	Status
L2 virtual circuit between two edge ports or VLANs	L2-SDX	Functional	Must	Completed
MPLS encapsulation of L2-SDX circuits	L2-SDX	Functional	Must	Completed[1]
VLAN and Stacked-VLANs (802.3ad) encapsulation	L2-SDX	Functional	Must	Completed[2]
IP transport between BGP peers	L3-SDX	Functional	Must	Completed
Custom route selection process	L3-SDX	Functional	Desirable	Planned
IPv6 support	Both	Functional	Must	Completed
Control plane resiliency	Both	Non-functional	Must	Completed
Control plane failure recovery	Both	Non-functional	Must	Completed
Network status after control plane failure	Both	Non-functional	Must	Completed
BGP control plane resiliency	L3-SDX	Non-functional	Must	Completed
Traffic rerouting after data plane failures	Both	Non-functional	Must	Completed
Control BGP attributes for each BGP peer	L3-SDX	Functional	Desirable	Not needed[3]
Apply separate policies for each BGP peer	L3-SDX	Functional	Desirable	Not needed[3]
Add, remove or shutdown BGP peers without impacts	L3-SDX	Functional	Must	Completed
Scale up to 100 BGP peers	L3-SDX	Non-functional	Desirable	Planned
Scale up to 100K routes	L3-SDX	Non-functional	Must	Planned[4]
Data plane statistic collection	Both	Operational	Desirable	Completed
Export of statistics to standard NMSs (SNMP, IPFIX)	Both	Operational	Optional	Ongoing[5]
Logging facilities	Both	Operational	Must	Completed

[1] Not fully supported in all switches [2] Stacked-VLANs not fully supported in the switches [3] Realized in the BGP peer [4] Tested up to 15K routes [5] SNMP not supported

Table 1. GEANT SDX requirements.

The SB layer consists of a collection of software modules called "providers," which interact with data plane devices using different southbound protocols. Providers gather information about network state and pass it to the distributed core, and receive instructions from the core to program the devices.

On the northbound side, ONOS presents abstractions to the applications, including Network Topology, Flow Objectives, and Intents. Intents provide applications with a network-centric programming abstraction that allows developers to program the network through the usage of high-level policies that capture *what* needs to be done, rather than how to do it. The Intent framework determines how to implement an intent based on what other policies are in the system, and abstracts low-level details of this implementation. Intents make network policy configuration easier, speed up management procedures, and tend to reduce the occurrence of configuration errors. Intents are backed by a dedicated subsystem that:

- Translates Intents into device instructions
- Coordinates and ensures the installation of the generated instructions
- Reacts to network changes and modifies paths accordingly
- Permits optimization across intents' translations

The Intent framework has been widely used for developing the L3-SDX and L2-SDX applications.

ONOS is supported by an active open source community. Different ONOS applications have been developed over the years by ON.Lab and by the community as listed in the documentation of the project. For example, the SDN-IP application allows SDN islands to seamlessly interconnect with external networks using standard BGP. Among the applications, we mention CORD™ (Central Office Re-Architected as Datacenter) [12]. It aims to revolutionize the way service provider central offices are built and operated. It brings in the principles of SDN, network functions virtualization (NFV), cloud technologies, and disaggregation, thereby making the central offices more manageable and agile.

High-Level Architecture for GÉANT SDX

The proposed SDX architecture is based on SDN enabled networking equipment, controlled by a cluster of ONOS controllers. The ISP services, such as L3-SDX and L2-SDX, are designed as northbound applications running simultaneously on top of the NOS, offering both L2 and L3 connectivity services. Coexistence of differ-

ent services in the data plane can be enforced through slicing mechanisms (e.g., VLAN tagging). As for the networking equipment, their integration is possible through open APIs, like OpenFlow, or by vendor-specific APIs implemented in ONOS in the form of pluggable drivers. The use of so-called *white box* devices is currently under investigation and testing in GÉANT as they could replace traditional equipment to achieve relevant cost savings.

An SDX can span a single location (e.g., replacing an OXP or PoP) or multiple locations (e.g., federating PoPs in a single logical PoP, or creating a distributed OXP). This issue is further discussed in the section about the practical experience. The L3-SDX has been developed on top of an existing ONOS application, called SDN-IP, while L2-SDX has been realized as a new ONOS application.

L3-SDX/SDN-IP in GÉANT

SDN-enabled networks still need to interoperate with traditional networks on the Internet. The ONOS SDN-IP application interconnects an SDN island with external networks leveraging the BGP protocol. The solution allows:

- External ASs to exchange routes and transit traffic through an SDN network
- The SDN network to advertise routes to the external networks
- A service provider to scale its SDN control plane by segmenting an AS into multiple SDN domains, which communicate through BGP

Besides the technical advantages, the service providers also gain benefits in reduced capital expenditures (CAPEX) and operational expenditures (OPEX), since they can use a single set of devices to manage L2 and L3 connectivity (and possibly L0/L1).

The high-level architecture of SDN-IP is shown in Fig. 4. The SDN network is composed of different data plane devices controlled by ONOS, which are directly connected to the BGP-speaking border routers of the external ASs. Finally, one or more internal BGP speakers peer with the external routers and act as bridges between the external domains and the SDN-IP application. From the legacy networks' perspective, the SDN domain appears as a standalone AS, as though it was running legacy BGP routers at the edges. Within the SDN network, SDN-IP has two main roles. The first is to install flows for BGP traffic between the external routers and the internal BGP speakers, thus allowing BGP sessions to be established. The second is to translate received routes into ONOS Intents, which are compiled down into flows on the SDN switches. In order to transport the data traffic in the SDN network, ONOS makes use of *multipoint to single-point* tunnels, avoiding the use of $n \times n - 1$ tunnels to connect the endpoints, thus reducing the flow table entries in the data plane.

SDN-IP provides a feasible migration path toward the softwarization of ISP networks. It can be integrated with networks that already use BGP both externally and internally. From an operational point of view, SDN-IP guarantees flexibility in the covered use case, as it does not make any assumptions on the deployment scenario. The application can run on one or multiple ONOS instances. Moreover, the BGP settings can be

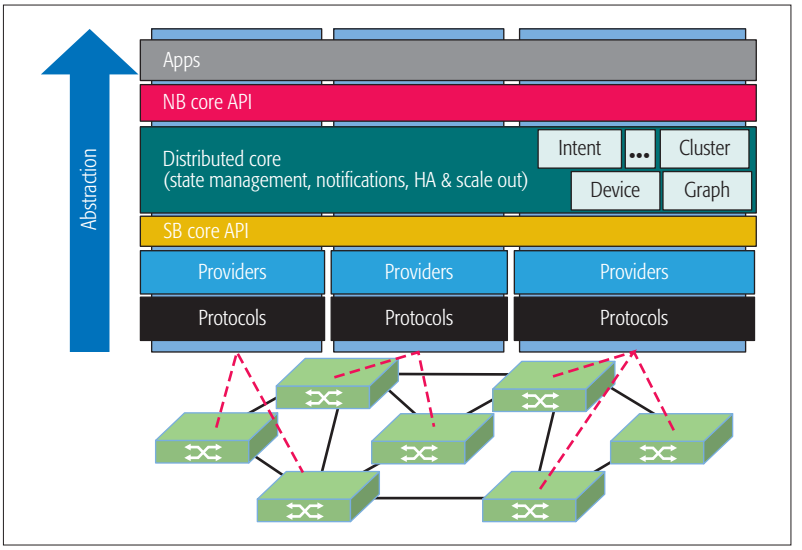


Figure 3. ONOS architecture.

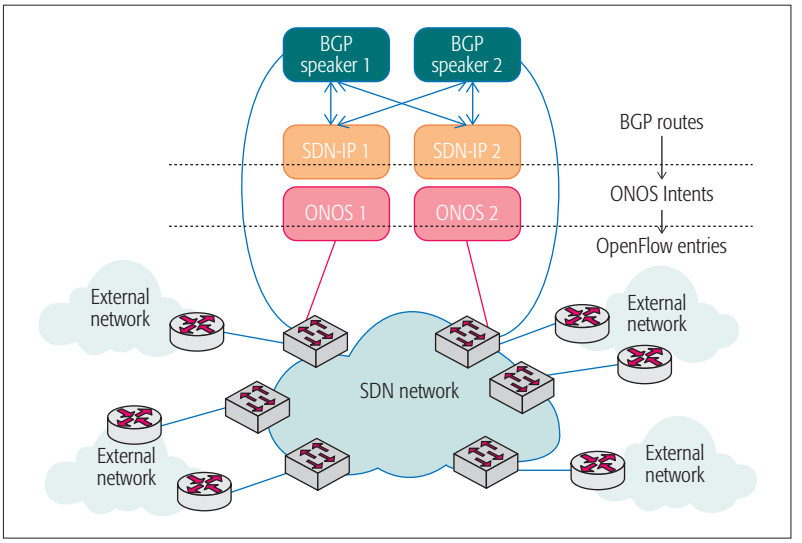


Figure 4. High-level representation of the SDN-IP architecture.

changed dynamically with the addition or removal of peers. SDN-IP provides high availability (HA) within the application itself: the service keeps working seamlessly, as long as there is at least one instance of the SDN-IP application running. In addition, SDN-IP leverages the HA mechanisms provided by ONOS for maintaining a consistent forwarding state in the data plane. SDN-IP provides a scalable solution able to control large-scale of SDN networks by using BGP-based confederations and ONOS clusters of different sizes.

L3-SDX extends SDN-IP, adding support for new deployment scenarios and providing facilities to monitor the BGP and transit traffic in the network. L3-SDX improves the flexibility of SDN-IP, making it possible to deploy multiple peers belonging to the same AS and interconnected through different connection points controlled by ONOS. The application supports the typical IXP scenario where all the BGP routers as well as the route server belong to the same subnet [6]. An integration with the ONOS IPFIX application allows exporting the counters related to the BGP sessions and to the L3 tunnels using the standardized IPFIX protocol. This can be used to realize

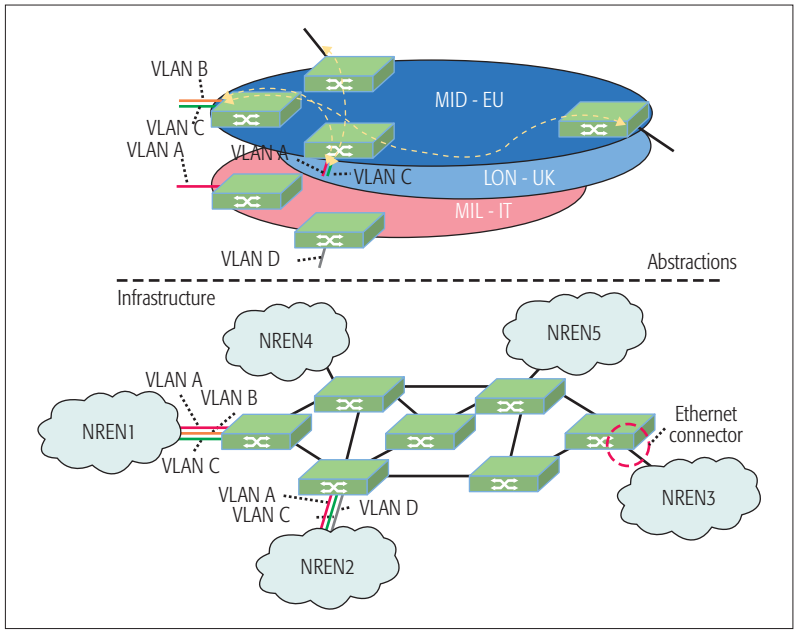


Figure 5. L2-SDX application and its abstractions.

advanced monitoring tools. L3-SDX and SDN-IP are both available under a liberal open source license.

L2-SDX Service in GÉANT

L2-SDX is an ONOS application that allows the automated provisioning of L2 tunnels between endpoints, which can be physical Ethernet interfaces or VLANs. The offered L2 services belong to the class of IP virtual leased line services (IP VLL) or virtual private LAN services (VPLS), which are a fundamental part of the service portfolio offered by large-scale ISPs. At the time of writing, only IP VLL has been integrated in the L2-SDX application, but VPLS can be provided with a straightforward extension of the current implementation. From a customer perspective, the L2-SDX appears as a black box that transports traffic from the source to the destination endpoint, as if they were in the same Ethernet LAN. Inside the SDX infrastructure, the L2-SDX application provides the necessary mechanisms for the service provisioning and monitoring. The human operators can manage and monitor the application through a command line interface (CLI) and a graphical user interface (GUI), which accepts high-level customer requests and translates them into ONOS *point-to-point intents*.

The application is fully integrated in ONOS and implemented as a callable service. In the next release, its services will be exposed through a REST API, allowing integration with orchestration platforms. Monitoring is achieved through the ONOS IPFIX application. L2-SDX can run over a single ONOS instance or on a cluster of ONOS instances that share common state information. The multi-instance deployment is useful to control large-scale SDXs made up of many SDN devices. L2-SDX leverages the high availability mechanisms provided by ONOS in order to maintain a consistent state in both the control and data planes. Failures in the control plane are managed through the redundancy of the ONOS cluster. Instead, data plane failures are automatically resolved through transparent re-routing mechanisms provided by ONOS.

L2-SDX provides users with powerful APIs and abstractions as shown in Fig. 5. A virtual SDX (e.g., MID-EU, LON-UK in the figure) contains a number of endpoints modeled as edge connectors, which can be interconnected through virtual circuits. Customers manage only the edges of the SDN network controlled by ONOS. The L2-SDX application eases service management and provisioning (e.g., enforcing isolation and avoiding several types of conflicts):

- The resources (ports or VLAN tags) associated with a connector cannot be reused.
- An edge connector can only be used in a single circuit.
- A connector in a virtual SDX instance cannot be interconnected with a connector in another virtual SDX.

L2-SDX is available under a liberal open source license.

Proofs of Concept and Worldwide Experimental Deployment

We realized two proofs of concept (PoCs). The first PoC was deployed in a laboratory at the University of Rome Tor Vergata and used mainly for validation and testing. It is based on virtual machines (VMs) and Open vSwitch (nine VMs emulate the data plane, a cluster of three VMs makes up the ONOS control plane).

The second PoC was realized in the GÉANT Testbed Service (GTS) [8]. The GTS delivers virtual testbeds powered by several facilities, co-located with GÉANT PoPs, offering different types of resources like VMs, SDN devices, and virtual circuits. Using GTS, we have built a large-scale PoC with seven HP OpenFlow switches deployed in seven PoPs (Fig. 6). This data plane is controlled by a cluster of three ONOS instances located in three of the PoPs. Three VMs, working as BGP peers, and two stub networks with perfSONAR hosts have been deployed. PerfSONAR is a performance measurement and troubleshooting tool for multi-domain scenarios.

The SDX PoC on GTS was integrated into a worldwide demo hosted at the Open Networking Summit 2016, where ON.Lab successfully deployed ONOS and SDN-IP, creating a global network facility entirely based on SDN. The network spans over 5 continents, interconnecting 9 RENs and more than 30 universities and research centers

Deployment Experience and Benefits for GÉANT

Overall, the SDX PoC on GTS worked according to our expectations; L2-SDX and L3-SDX passed all the functional tests we performed. The *status* column in Table 1 reports the coverage assessment of the input requirements. The scalability and efficiency of L2-SDX and L3-SDX are tightly related to ONOS performance, and it has been demonstrated in [7] that the platform can meet carrier grade requirements in specific deployment conditions. L3-SDX and SDN-IP can scale up to 15,000 routes, achieving the current GÉANT requirement of 12,000 announced routes.

The SDX deployed in the GTS represents a single geographically distributed SDX, spanning seven PoPs, with three ONOS instances running in different countries. During the execution of the functional tests, we gained feedback (e.g., the mastership election duration) that drove us not to further stress the SDX under critical events like controller instance failure or data plane failures. Therefore, we believe that having single-location SDXs, spanning a single OXP or PoP, is a safer approach to start the migration toward SDN. OXPs are good candidates for early deployment of SDX due to the complexity of the services that are offered, which makes the introduction of the SDN-based approach attractive. Moreover, with single-location SDXs, the devices can keep their independence and troubleshooting capabilities. In the current GÉANT network, each PoP is seen as a "hop" by the IP traffic, while in a geographically distributed L3-SDX, simple troubleshooting tools like traceroute would no longer be useful. Incidentally, we observe that there is a gap to be filled with troubleshooting tools for SDN-based networks, because L3 tools based on ICMP (ping, traceroute) do not work hop by hop in a network of SDN controlled switches.

The transition toward geographically distributed SDX could start with federations of nodes controlled by the same NOS instance. We have done some preliminary work on this issue with ICONA [14], an application to interconnect multiple ONOS clusters seamlessly through an "East-West" interface. Initially, OXPs could be interconnected, creating a geographically distributed L2 fabric controlled by ONOS. Likewise, geographically close PoPs could be federated in small clusters.

From the point of view of the development costs, it was possible to release the L2-SDX and L3-SDX in the PoCs with relatively little effort (on the order of three man months) thanks to the possibility of relying on the ONOS code base and documentation.

Let us now provide a high-level analysis of the benefits achievable by GÉANT with the softwarization of the infrastructures, in terms of operational costs like services provisioning and services management. The deployment of an SDX in place of an OXP will automate most of the configuration operations, reducing the efforts of the human operators. Currently, in order to set up a GÉANT OPEN service between two access points (ports or VLANs) inside an OXP, the customer has to contact the operators who manually configure the connection [8]. These operations (creation of virtual interfaces, VLAN id selection on both endpoints, VLAN id rewriting) are error-prone and require coordination between the interested parties. Any arising issue requires further manual intervention of the operators. A typical target for the provisioning time of these services is five days. Using the L2-SDX, it will last minutes instead [15]. Moreover, most failure cases are automatically resolved by L3-SDX/L2-SDX using ONOS built-in mechanisms (e.g., a failure of a controller is solved using redundancy of controller instances, a switch failure is solved by ONOS with re-computation of data plane paths around the faulty switch).

As regards GÉANT IP and GÉANT Plus, similar improvements in the services provisioning and management are an affordable objective with single-location SDX. They represent a tangible result when compared to the current procedures (five days to obtain IP connectivity or to set up an L2 circuit).

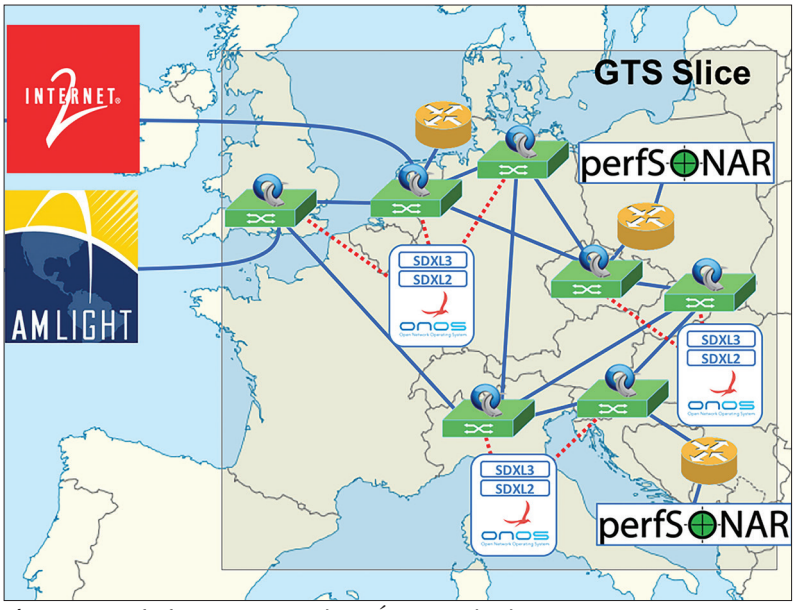


Figure 6. Proof of concept over the GÉANT testbed service.

When considering geographically distributed SDXs, having a centralized view of the network potentially brings further benefits like more efficient traffic management, but the challenges for a wide area SDN deployment still need to be solved. In particular, a first issue is the impact of the latency and unreliability of the control plane connections between controllers and remote network nodes. A second issue arises when the controller instances are distributed in a geographical way, in order to reduce the latency toward the controlled nodes. The mechanisms used to achieve a consistent view across all the distributed controllers works well in LANs, where the latency is low and the capacity is high. The performance of these consistency mechanisms can become critical in geographically distributed wide area networks; a careful assessment of these aspects is still needed.

Finally, let us consider an NREN that would like to establish L2 connectivity with a third party (not GÉANT). Different configurations have to be in place in order to have the connectivity operational:

- L2 configuration of the PoP where the NREN is located
- L2 configuration of the PoP where the target OXP is located
- Steering of the NREN flow into a label switched path (LSP), provided by MPLS/BGP, toward the destination PoP
- Establishment of L2 connectivity inside the OXP

Despite the services being logically similar, the provisioning procedure can require up to 10 days, because they are managed in two completely separated infrastructures, with specific hardware, and different policies and configuration mechanisms. Moreover, coordination is needed between all the operators (GÉANT, NREN, and

In order to bring the L2-SDX and L3-SDX services to production, some additional concerns related, for example, to security and integration with management systems need to be addressed. The GÉANT team is committed to working on these issues in the near future.

the third party). Instead, within an SDX environment, the separation is blurred, and these services can be managed through a single platform, reducing coordination efforts, manual intervention of the network operators, and complexity of the procedures. SDX allows a network engineer to provision an L2 circuit without prior technical coordination with the other network teams [15]. Moreover, they are delivered using a unique infrastructure, with reductions in terms of OPEX and future hardware investments.

Conclusions

In this article, we have first considered the long-term softwarization path of a service provider network like GÉANT. Then we have described the prototypes of two services, called L3-SDX and L2-SDX, that have been deployed in a proof of concept over the GÉANT Testbed Service. The development has been based on the open source ONOS controller platform for SDN. We have performed a functional evaluation of the PoC and an analysis of the potential benefits, which have been very satisfactory. The developed services can bring considerable savings in the operational costs and can dramatically reduce the service provisioning time, as they automate many tasks that are manually performed. From the point of view of performance, the SDX solution is ready for a "local" deployment, that is, considering a single location (even if composed of a large number of nodes). It has to be further assessed if geographically distributed locations can be combined in single logical instances of the SDX.

In order to bring the L2-SDX and L3-SDX services to production, some additional concerns related, for example, to security and integration with management systems need to be addressed. The GÉANT team is committed to working on these issues in the near future.

Acknowledgments

This work has been partially funded by the EC under project GN4-1.

References

[1] Open Networking Foundation, "Software-Defined Networking: The New Norm for Networks," ONF White Paper, Apr. 13, 2012.
[2] A. Galis *et al.*, "Softwarization of Future Networks and Services-Programmable Enabled Networks as Next Generation Software Defined Networks," *IEEE Wksp. SDN for Future Networks and Services*, Trento, Italy, Nov. 2013.
[3] M. Kind *et al.*, "Softwarization of Carrier Networks," *Info. Technology*, vol. 57, no. 5, Oct. 2015, pp. 277–84, ISSN (Online) 2196-7032, ISSN (Print) 1611-2776, DOI: 10.1515/itit-2015-0019.
[4] J. Mambretti *et al.*, "Software-Defined Network Exchanges (SDXs) and Infrastructure (SDI): Emerging Innovations in SDN and SDI Interdomain Multi-Layer Services and Capabilities," *2014 1st Int'l. IEEE Science and Technology Conf. (Modern Networking Technologies)*, 2014, pp. 1–6.
[5] I. Monga, "Software Defined Exchanges: The New SDN?," http://meetings.internet2.edu/media/medialibrary/2015/03/31/Monga_SDX_SDN.pdf, accessed Dec. 30, 2015.
[6] I. Pepelnjak — "Could IXPs Use OpenFlow To Scale?", http://www.menog.org/presentations/menog-12/116-Could_IXPs_Use_OpenFlow_To_Scale.pdf, accessed Dec. 30, 2015.
[7] ONOS Project — ONOS homepage, http://onosproject.org, accessed Jan. 5, 2016.
[8] GÉANT Connectivity and Network Management Services — Connectivity and Netwok Management Services home page, http://www.geant.org/Services/Connectivity_and_network/Pages/Home.aspx, accessed Feb. 5, 2016.
[9] A. Gupta *et al.*, "Sdx: A Software Defined Internet Exchange," *Proc. 2014 ACM Conf. SIGCOMM*, 2014, pp. 551–62.
[10] J. Chung *et al.*, "AtlanticWave-SDX: An International SDX to Support Science Data Applications," 2015.
[11] R. Buyya *et al.*, "Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering It Services as Computing Utilities," *10th IEEE Int'l .Conf. High Performance Computing and Commun.*, 2008.
[12] CORD Project — CORD homepage, http://opencord.org. Accessed Jan 15, 2016.
[13] B. Claise (Ed.), B. Trammell (Ed.), and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information," IETF RFC 7011, Sept. 2013.
[14] M. Gerola *et al.*, "ICONA: Inter Cluster ONOS Network Application" *2015 1st IEEE Conf. Network Softwarization*, 2015, pp. 1–2.
[15] J. Ibarra *et al.*, "Benefits Brought by the Use of OpenFlow/SDN on the AmLight Intercontinental Research and Education Network," *2015 IFIP/IEEE Int'l. Symp. Integrated Network Management*, 2015, pp. 942–47.

Biographies

Pier Luigi Ventre (pier.luigi.ventre@uniroma2.it) received his Master's degree in computer engineering from the University of Rome Tor Vergata in 2014, with a thesis on *Information-Centric Networking and Software Defined Networking*. From 2013 to 2015, he was a beneficiary of the Orio Carlini scholarship granted by the Italian NREN GARR. His main research interests focus on computer networks, software defined networking, virtualization, and information-centric networking. He has worked on several EU research projects, and currently he is a Ph.D. student in electronic engineering at the University of Rome Tor Vergata.

Stefano Salsano [SM] (stefano.salsano@uniroma2.it) received his Ph.D. from the University of Rome "La Sapienza" in 1998. He is an associate professor at the University of Rome Tor Vergata. which he joined in 2000 as an assistant professor. He has participated in 15 research projects funded by the EU, being technical coordinator of two of them. He has been a principal investigator in several research and technology transfer contracts funded by industries. His current research interests include software defined networking, network virtualization, cybersecurity, and mobile and pervasive computing. He is a co-author of an IETF RFC and more than 130 peer-reviewed papers and book chapters.

Matteo Gerola (mgerola@fbk.eu) is a software architect and senior research engineer at the Future Networks unit at Bruno Kessler Foundation (FBK) CREATE-NET research center. His main research interests focus on SDN, network virtualization, OpenFlow, and Optical Networks. Within FBK CREATE-NET he has been involved in several European projects on SDN, optical technologies, and future Internet testbeds. He has published in more than 20 International refereed journals and conferences. Over the years, he has participated in events, conferences, and workshops as TPC member, author, and invited speaker.

Elio Salvadori (esalvadori@fbk.eu) received his M. Sc. degree (Laurea) in telecommunications engineering from Politecnico di Milano in 1997 and then worked as a network planner and systems engineer in Nokia Networks and Lucent Technologies until 2001, when he moved to the University of Trento to receive his Ph.D. in 2005. He is currently acting as director of CREATE-NET, a research center focused on telecommunications and computer networks, part of FB) based in Trento. His main research interest are software-defined networks and optical networking.

Mian Usman (mian.usman@geant.org) Mian is the Network Architect at GÉANT, Mian received his BSc in Network Management and Design from University of Portsmouth in 2007 and MBA from Manchester Business School in 2017. Mian's work is focused on network architecture and design he led the technical IP team responsible for designing and deploying GEANT's new IP/MPLS platform and the migration of EoSDH services to EoMPLS.

Sebastiano Buscaglione (sebastiano.buscaglione@geant.org) is a professional in the field of networking with several years of experience working in large-scale service provider networks. Before joining DANTE, now GÉANT, in 2012 where he is currently employed as a senior network engineer, he worked as part of the AT&T global operations department supporting global enterprise VPN services. His main interests revolve around extraction and analysis of network data and its use in driving optimization in network architectures. His study includes net-

working at the Cisco Networking Academy at Metropolitan University, London, United Kingdom, and industry certifications such as CCNP and MEF-CECP.

Luca Prete (luca@onlab.us) is an SDN enthusiast, currently leading the SDN deployment activities at Open Networking Laboratory (ON.Lab). He received his Bachelor's degree in computer science from the University of Milano and his Master's degree in Internet technologies from the University of Pisa. He has been involved in computer science consulting since 2005. Before joining ON.Lab, he collaborated with the Italian NREN in the R&D Department, focusing his attention on software defined networks.

Jonathan Hart (jono@onlab.us) received his Bachelor of Engineering with Honours in network engineering from Victoria University of Wellington, New Zealand, in 2011. He has worked as a software engineer on SDN projects at ONLab in California for the past four years, and is currently a core team member of the ONOS and CORD open source projects.

William Snow (bill@onlab.us) is the chief development officer with ON.Lab. He is responsible for all engineering and operations at ON.Lab and leads the teams providing core engineering to the ONOS and CORD projects. Prior to joining ON.Lab, he spent over 25 years in the industry building development teams and delivering innovative products. He has led engineering teams for both startups and public companies in the networking and security spaces. He was responsible for the routing and high availability teams delivering the Cisco CRS-1. He was also responsible for the Centillion LAN switching product line prior to Centillion's acquisition by Bay Networks. He received his Bachelor of Science in electrical engineering from Cornell University, a Master of Science in electrical and computer engineering from Stanford University, as well as a Master of Science in engineering management from Stanford University.

Manufactured by Software: SDN-Enabled Multi-Operator Composite Services with the 5G Exchange

Gergely Biczók, Manos Dramitinos, Laszlo Toka, Poul E. Heegaard, and Håkon Lønsethagen

The authors introduce the 5G Exchange (5GEx) concept that builds on SDN and NFV, and facilitates the provisioning of multi-operator 5G services by means of inter-operator management and orchestration of virtualized network, compute, and storage resources.

Abstract

Foreseen 5G verticals hold the promise of being true value-added services, hence bringing significant income to their respective providers. However, the nature of these verticals are very demanding in terms of both economic and technical requirements, such as multi-operator cooperation, end-to-end quality assurance, and the unified orchestration of network and cloud resources. Existing systems fall short of satisfying these requirements, but emerging network softwarization and resource virtualization technologies, such as SDN and NFV, show promise for being key enablers in this context. In this article, we introduce the 5G Exchange (5GEx) concept that builds on SDN and NFV, and facilitates the provisioning of multi-operator 5G services by means of inter-operator management and orchestration of virtualized network, compute, and storage resources. We present potential 5GEx use cases, conceptual architecture, and value proposition. We also outline open research questions on how to exchange information in such a coopetitive environment, and provide an outlook on the impact of 5GEx on a network service provider's business and operation.

Introduction

Internet services have evolved rapidly, covering all aspects of communication and infotainment. Services such as video on demand and online gaming are already popular, while additional verticals, also integrating cloud and the Internet of Things (IoT), are envisioned in the context of the fifth generation (5G) [1]. Efficient provisioning of these services as high-value products in the market requires service-aware routing, *end-to-end* quality of service (QoS) assurance, including dependability aspects, elastic resource, and dynamic service orchestration (over network and cloud infrastructures), and flexible service management. These requirements are currently not met by best effort Internet and the inherent shortcomings of a single-traffic-class approach: large buffers for statistical multiplexing gain inevitably increase delay; there is no way to protect critical over non-critical traffic; the flow control protocols cannot efficiently adapt to congestion and match application requirements with network capabilities, while Border Gateway Protocol (BGP) does not allow multiple choices for service-aware routing of delay-tolerant vs. delay-critical traffic so as to both optimize QoS and load balance the network.

Internet service layer stakeholders buy and sell Internet services and are categorized as connectivity providers, information providers, also referred to as over the top (OTTs), and end users. Connectivity providers, also referred to as network service providers (NSPs), normally own their network and are responsible for the provisioning of its functionalities. The fact that multiple network (including 5G radio access), cloud, and OTT stakeholders constitute the multi-actor value chain of 5G services inevitably calls for *multi-operator business and service coordination jointly over the network, compute, and storage domain*. Unfortunately, currently there is no open and global solution to multi-service internetworking, resulting in costly, legacy, provider-specific service provisioning. This limits the potential of standardized integration of network, compute, and storage infrastructure under a *unified* service orchestration, control, and management framework. Thus, it is insufficient to carry the 5G value creation at the edge of the networks across the backbones, hindering the 5G services value creation. The softwarization of the network control plane and the virtualization of resources can be powerful enablers in the context of a novel exchange mechanism supporting on-demand service creation, standard resource abstractions, resource trading, and flexible inter-provider service level agreements (SLAs). Such an exchange framework has the potential to remove the inherent shortcomings of today's solutions, and enables 5G value creation at the network edge and matching of requirements of 5G applications and services to properties of connectivity services end to end over the virtualized network infrastructure.

Multi-operator services currently rely on best effort connectivity enabled by service-agnostic interconnection agreements pertaining to inter-domain traffic aggregates. As a result, services experience unpredictable network performance that mostly depends on insufficient and inefficient overprovisioning [2]. Paired with the increasing overall traffic demand and the limited

Gergely Biczók is with Budapest University of Technology and Economics; Manos Dramitinos is with Athens University of Economics and Business; Laszlo Toka is with MTA-BME Information Systems Research Group; Poul E. Heegaard is with the Norwegian University of Science and Technology; Håkon Lønsethagen is with Telenor.

Digital Object Identifier: 10.1109/MCOM.2017.1600197

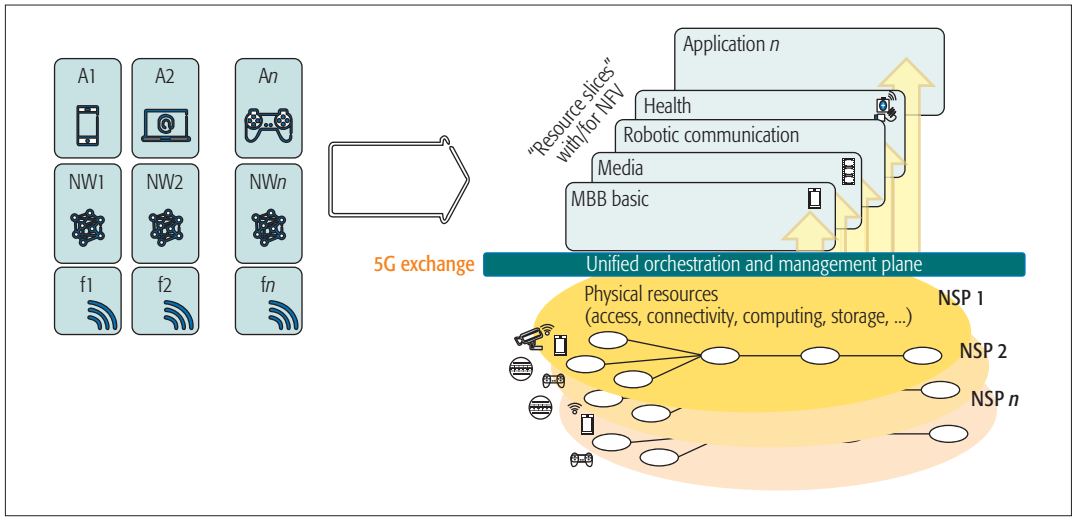


Figure 1. From dedicated physical resources to network factory: services manufactured by software.

The ETICS research project has attempted to mitigate these inefficiencies by complementing traditional interconnection peering and transit products with additional products for the provisioning of quality assurance to the inter-domain interconnection services termed Assured Quality products.

incentives for investing in new infrastructure, 5G services provisioning is a formidable challenge. Overprovisioning is a bad strategy for parts of the network, since a capacity upgrade also brings substantial benefit to a less well provisioned but interconnected network, making the latter also more attractive to end users [3]. Thus, although both networks could benefit from an upgrade, selfishly maintaining a low-quality interconnection is often a dominant strategy for large Internet service providers (ISPs) [4].

In this article, we introduce the novel 5G Exchange (5GEx) concept: drawing on the disruptive innovative technologies of softwarized service orchestration and control (SDN) and resource and network function virtualization (NFV), 5GEx is positioned to be a key market enabler for the provisioning of multi-operator infrastructure services and a catalyst for materializing the value of 5G verticals. From the technical resource orchestration aspect, 5GEx enables the transition from dedicated physical networks and resources for different applications to a "network factory," where resources and network functions are traded on demand and new services are "manufactured by software" (Fig. 1). Orchestration of the heterogeneous resource domains are achieved via virtualization of resources and network functions and a smart slicing method operating over those virtualized entities.

The rest of the article is organized as follows. We present the state of the art in multi-operator interconnection alternatives. We present the benefits of SDN and NFV, and how these characteristics enable the proposed exchange framework. We introduce the conceptual design of the 5G Exchange. We discuss operational challenges and opportunities in the context of the exchange from the operator's aspect. Finally, we conclude the article.

Multi-Operator Interconnection: State of the Art

Multi-operator services have been implicitly supported over the Internet by means of pure connectivity services and interconnection agreements between the operators. Networks rely on BGP to build the Internet connectivity graph. They solely exchange BGP announcements and data; interconnection agreements specify whether and how each network should accept and terminate or forward the traffic coming from a neighboring network. Existing *peering* and *transit* interconnection agreements do not provide any type of service-aware routing and management or QoS assurance, and pertain to inter-domain traffic aggregates of multiple services (elastic and inelastic).

From an economic standpoint, the work in [3] points out inefficiencies, such as unfair revenue distribution among providers, which discourages network upgrades. The current "walled garden" digital service provisioning regime (i.e., operators focusing on intra-domain services for their customer base while being reluctant to cooperate with other operators beyond peering and transit) results in well-known business and economic inefficiencies. This motivates an increasing research and business interest in providing solutions for enabling sustainable ecosystems where services relying on open, agile, elastic management, as well as *quality assurance* can be efficiently provisioned [2].

The ETICS research project has attempted to mitigate these inefficiencies by complementing traditional interconnection peering and transit products with additional products for the provisioning of quality assurance to the inter-domain interconnection services, or assured quality (ASQ) products [5]. ASQ products support technology-agnostic paths of assured performance and attributes such as IP addresses/prefixes, delay, jitter, and bandwidth; service level agreement (SLA) attributes are propagated via an overlay of path communication elements, horizontal and vertical interfaces that map them to and enforce them in the underlying networks. ASQ products allow a finer degree of traffic control over inter-domain network paths and regions, while peering pertains to two networks and transit offers global connectivity to the buyer.

Evolving from a separate industry strand, the IP Exchange (IPX) is a non-Internet telecommunications interconnection model developed by the GSM Association (GSMA) for the exchange of IP-based traffic between customers of sepa-

While there has been tremendous activity in the networking research community with regard to service quality assurance spanning multiple decades, we believe that the time has just become ripe for capitalizing on a well-designed solution.

rate mobile and fixed operators as well as ISPs; essentially, a privately managed IP backbone [6]. IPX, in contrast to the Internet approach, inherently supports QoS interconnection with respective SLAs and cascading payments between operators. There are two characteristics of IPX that work against IPX becoming a truly global solution. First, IPX uses network address translators and per session gating at IPX border routers, making it hard to scale at a global level. Second, IPX is required to be compatible with legacy voice services, adding additional complexity. These characteristics render IPX too complex and expensive for serving as a general-purpose backbone for 5G services.

From a cloud provider's standpoint, **cloud federations and exchanges** are gaining momentum, including data center interconnection over federated infrastructure. A typical example is OnApp [7], a federation of cloud providers with cloud and content distribution network (CDN) product offerings of fine geospatial granularity worldwide. In addition, U.K.-based CloudStore [8] supports public, private, and hybrid clouds, and Inernet as a service (IaaS), software as a service (SaaS), platform as a service (PaaS), and specialist cloud services. While certainly a hotbed of technical and business innovation recently, solutions from the cloud domain have to be complemented and harmonized with end-to-end assured quality connectivity to successfully provision 5G services.

Enabling Technologies: SDN, NFV, and SFC

While there has been tremendous activity in the networking research community with regard to service quality assurance spanning multiple decades, we believe that the time has just become ripe for capitalizing on a well designed solution. Two of the key requirements have just materialized:

1. We have the "killer apps" in form of value-added 5G verticals whose value chain is inherently multi-operator in general.
2. Enabling technology has just reached the needed maturity level, in the form of software-defined networking (SDN), network functions virtualization (NFV), and service function chaining (SFC), and how these technologies can be integrated around the concept of network (and later compute and storage resource) slices.

SDN, in its original interpretation, decouples control from the data plane (and therefore vendor-specific hardware), assigning it to a software controller. SDN simplifies routers and switches, and can improve data throughput and reduce congestion via traffic management and optimized resource allocation applied by the controller. In the context of 5G, SDN enables service-aware routing, flow-level quality assurance, and efficient dynamic resource management by a logically centralized control logic. The defining benefit of SDN for us, however, lies in its ability to provide an abstraction of the physical network infrastructure. SDN provides network programmability: several customized network slices can be configured in parallel using the same physical and logical infrastructure. Thus, one physical network can support a variety of services in an optimal way.

NFV allows for a network function to be implemented in software instead of by a piece of dedicated hardware. This concept comes with inherent scalability supporting the delivery of on-demand, dynamically re-scalable, and global services. For us, the key feature of NFV lies in its ability to execute NFs independent of location; this essentially means that the same NF can be executed at different locations for different network slices. Hence, a service-aware virtual network environment is created by the actual placement of NFs.

SFC is not a novel concept in itself: the delivery of end-to-end services often requires various network service (e.g., firewall) and application-specific functions (e.g., HTTP header processing) to be "chained." However, if functions are virtualized and can be placed at arbitrary physical locations, and SDN policies are used to steer data traffic through them in a service-specific manner, we have the ultimate elastic service environment: instantly, rapid creation, destruction, and scaling, and migration of service functions and (with an agile service insertion model) services become possible.

SDN and NFV allow architects to build systems with greater degrees of freedom and abstractions, and thus network flexibility: with their help the vertical networking of yesterday can be broken down to building blocks that can be chained together to suit the services to be supported. We refer to this concept as *service-aware slicing*; we believe it has the potential to enable the highly coveted flexibility in service provisioning and delivery (the "Holy Grail of 5G"), while reducing overall costs at the same time.

SDN is also a key enabler from a multi-operator collaboration perspective, as it allows for the direct expression of flexible policies potentially tailored to different applications and service quality requirements (a potential stepping stone for an improved SLA framework). Drawing from (a simplified version of) this idea, the project SDN eXchange point (SDX) [9] proposes to deploy SDN-capable switches at Internet exchange points (IXPs) in order to make a step from conventional hop-by-hop, destination-based forwarding and enable participating ISPs to apply diverse actions on packets from the IXP such as inbound traffic engineering, redirection of traffic to middleboxes, and load balancing. We believe that SDX is a step in the right direction, and serves as an important precursor to 5GEx.

The 5G Exchange Concept

The 5G Exchange project aims to enable 5G verticals by designing an exchange framework capable of handling the orchestration of both network and cloud resources over multiple technological and administrative domains [10]. Apart from catering to the needs of future 5G services, 5GEx also has an objective of overcoming the historical technological and market fragmentation of the European telecommunications sector by bootstrapping operator collaboration with regard to infrastructure services. Such infrastructure services (and associated resources) provide the foundation of all 5G verticals making use of cloud and networking services, apart from the radio interface itself. The envisioned 5G Exchange will enable operators to buy, sell, and integrate

virtual resources and services, thus enabling one-stop shopping for their customers: it suffices for the customer to contact and contract with a single operator, who will then outsource part of its commitments to other operators given the lack of geographical footprint or available resources. Furthermore, the generic, open, and standardized offering of various connectivity modes supported with other 5G capabilities will enable the numerous small to medium-sized enterprises (SMEs) and content providers to differentiate and monetize their online content and application offerings. This will open up new venues of innovation for many businesses and verticals as yet unseen in various consumer, business and public sector markets.

5G services extend personal communication and video services with the integration of cloud, IoT, and machine-to-machine communication into the 5G architecture and service model. The Next Generation Mobile Networks (NGMN) Alliance [1] specifies 24 use cases for 5G, to be delivered across various devices (smartphone, wearable, machine module) grouped into 8 families, along with related customer-facing services (verticals): 3 families of high-speed broadband access everywhere with HD video sharing as vertical, massive IoT with sensor/smart home networks as verticals, 3 families of lifeline/ultra-reliable communications with e-health/telemedicine as a vertical, and broadcast services for infotainment. These use case families and verticals motivate the wholesale infrastructure services needed to support them, enabled by resource virtualization, network softwarization. and service orchestration and management of the proposed 5GEx multi-operator exchange. 5GEx can be seen as the 5G evolution of exchange environments such as IPX, SDX, and IXPs; a core subset of 5G infrastructure services envisioned that are capable of supporting the aforementioned use cases and verticals are connectivity (e.g., VPN+), virtual network function as a service (VNFaaS, e.g., vCDN), and anything as-a service (XaaS, e.g., Gi-LAN).

Connectivity is a use case family of wholesale connectivity services over multiple domains, capable of supporting next-generation connectivity verticals such as VPN+. VPN+ denotes an improved virtual private network service (aimed at an enterprise customer) with a network as a service (NaaS) element such as partial topology description and dependability requirements. The vCDN use case implements a virtual CDN, where a video portal (customer) purchases the right to use the CDN facilities of a CDN provider: this also involves storage resources. XaaS represents the most challenging use case in that it potentially involves the full range of *network*, *compute*, and *storage* resources with strict performance and dependability guarantees (e.g., ultra-low latency and adequate computational capacity) to enable demanding verticals such as industrial robotics and mobile edge computing (MEC) scenarios. A very concrete instantiation of XaaS inspired by a true operator need is international mobile data roaming (referred to as the Gi-LAN use case). Since the European Union will ban mobile roaming fees within Europe, a drastic increase of roaming data usage is expected. The normal process of roaming would involve building tunnels back to the home packet gateway (h-PGW) through international exchange points: an expensive and unfavorably scaling mechanism. Moving the h-PGW and the entire Gi-LAN functionality to the roaming operator's data center results in a cheaper and dynamically scalable solution. (While this list of use cases is clearly not exhaustive, they demonstrate the different expected capabilities of 5G Exchange well.)

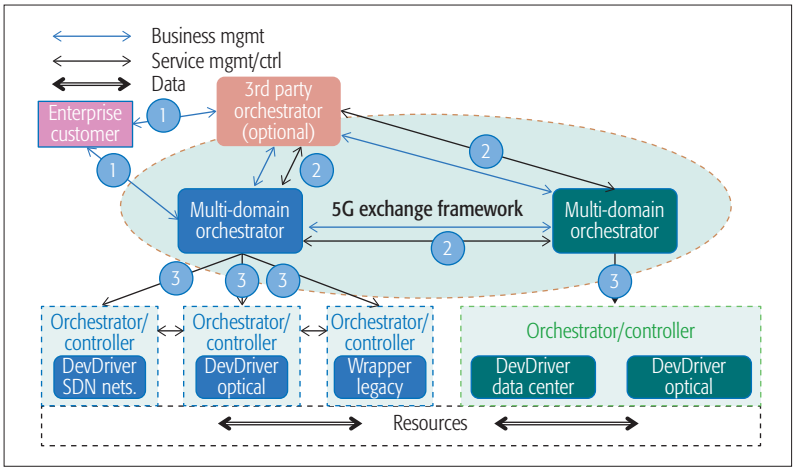


Figure 2. Simplified conceptual architecture of the 5G Exchange.

The aforementioned wholesale 5GEx infrastructure services can efficiently serve verticals by relying on lower-level 5GEx fundamental services and SDN/NFV techniques, with the core element being the *slice*. A slice is defined as a managed set of 5G resources and network functions set up within the 5G system that is tailored to support a particular type of user or service. Note how the slice concept is an evolution of network slices as used in SDN, also incorporating cloud resources and NFs. These slices are then instantiated on demand using application programming interfaces (APIs) exposed by the management plane, which provides dynamic orchestration for multilayer and multi-domain networks. SDN and NFV greatly simplify slice and service orchestration and management, as opposed to the traditional interworking of legacy networks and clouds. 5GEx uses:

- Standard interface (1) for the multi-domain orchestrator to translate the 5GEx customer service request to a chain of VNFs and underlying network, storage, and cloud resource requirements
- Standard interface (2) and respective SLAs for trading slices and 5GEx higher-level services among 5GEx-enabled orchestrators
- Standard interface (3) for the management of own or leased — via interface (2) – resources. For a simplified conceptual architecture of 5GEx, please refer to Fig. 2

Precursor projects containing ideas and code for interfaces include ETICS (interface 2, [5]), UNIFY (interface 3, [11]), and T-NOVA (interface 1, [12]).

The proposed 5G Exchange Framework supports *a variety of specific deployments and coordination/collaboration models* such as:

- "Direct peering" at an already established local or remote IXP
- Distributed multi-party collaboration, where the operators host the exchange mechanism in a distributed manner inside their own infrastructure

When such a managed and assured quality traffic exchange solution is enabled among a set of initial partners, the solution can scale and grow into a full fledge traffic, resource and service trading and exchange platform. This holistic multi-domain resource slicing solution will unleash the full potential of 5G.

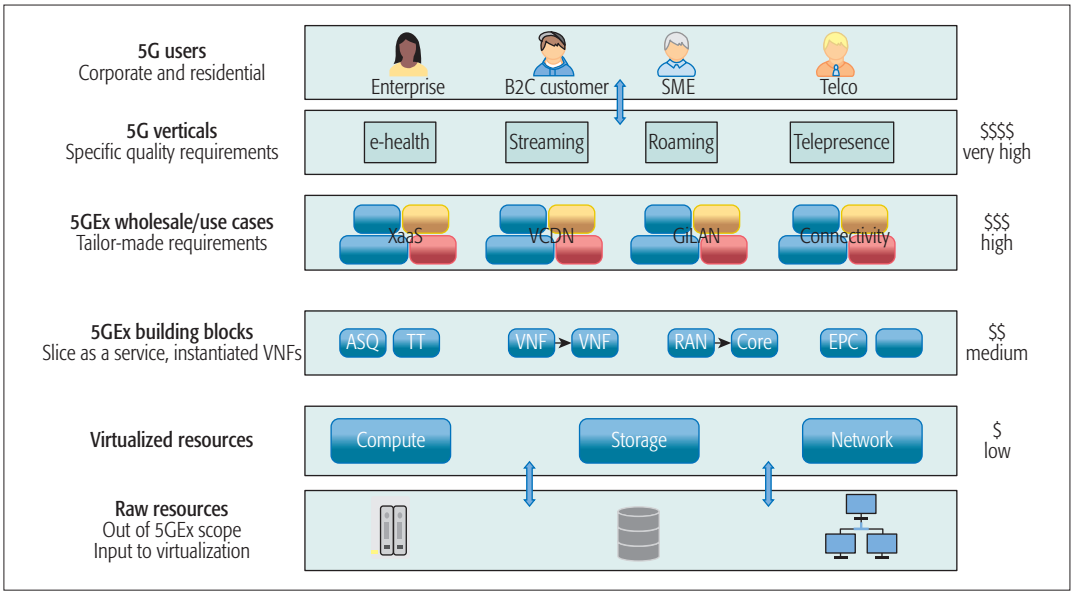


Figure 3. Levels of 5GEx goods and value proposition mapped to the current cloud ecosystem.

- A dedicated (for-profit) exchange point provider as a standalone entity, offering exchange point services

In addition, the 5G Exchange Framework also supports higher-level abstractions and advanced models covering views, resources, and services across several exchange points or points of presence (PoPs). Also note that the customer-facing "3rd party orchestrator" in Fig. 2 is an optional role in the ecosystem, essentially referring to a virtual network operator who implements the multi-domain orchestrator functionality, but does not own an infrastructure.

The clear separation of functionality allows 5GEx to make the most out of SDN and NFV, creating an open agile management and orchestration environment where multi-operator services become only marginally more complex than single-operator services due to the common interfaces and functionality for the management of both owned and leased services. There is an analogy here with the cloud ecosystem in terms of architecture and value proposition, with the various 5GEx layers [13] mapped to cloud resources and services ranging from Amazon's S3 through EC2 to CloudFront high-level service, as depicted in Fig. 3. Lower-level resources are the *low-margin commodity building blocks* of differentiated higher-level services for serving 5G verticals. At all levels of the value chain, we use the service concept appropriate for the given level. Virtual resources and NFs are composed into slices utilizing the network function infrastructure as a service (NFVIaaS) paradigm; slices make up infrastructure services in the concept of slice as a service (SlaaS); finally, infrastructure services comprise a custom 5G vertical, which in turn can be purchased by a customer residing outside the 5G Exchange (high profit margin). Therefore, intra-5GEx services by design consider the needs of 5G customer-facing services. The (potentially recursive) trading of slices and 5GEx services over interface (2) can support multiple coordination models, including push and pull, if goods are built on demand or pre-built and part of a catalog, as well as distributed or centralized, whether the multi-domain orchestration is a centralized third-party service or implemented in a distributed fashion over multiple instances, one per 5GEx infrastructure service provider.

Although the conceptual architecture and value proposition are clear, there are still some design challenges ahead, the most glaring focused on the exchange of information among 5GEx operators (see Fig. 4 for an example with User1 as a customer and Op1 as a customer-facing operator). On one hand, SDN supports a rich set of possibilities for exchanging information as described earlier. On the other hand, what kind of information should be shared, at what granularity, and how to calculate relevant key performance indicators (KPIs) and design corresponding SLAs are all mostly open questions. Naturally, there is a trade-off between the business interests (sharing the least amount of information possible) and optimal end-to-end resource management guaranteeing assured quality (sharing the most precise information possible). Furthermore, virtualization involves aggregation of information about the physical resources, network topology, and policies; multiple levels of virtualization in 5GEx make matters even more complex. Thus, the chosen method of information exchange has far reaching consequences with regard to both performance and dependability of the provisioned services.

In order to rise up to this challenge, the following research roadmap could be followed. First, we have to understand what is the maximum amount and finest granularity of information we could possibly collect using advanced SDN and cloud monitoring techniques. Second, we should carefully investigate and quantify the interdependence of virtual compute, storage, and network resources with regard to both performance and dependability, both within a domain and over multiple domains, or if they share some physical resources. Third, we should repeat step 2 upward in the value chain to have a model of the whole ecosystem. Finally, we should go beyond basic aggregated mean values when it comes to KPIs, and consider quantiles and even full probabili-

ty distributions of important performance and dependability metrics, potentially leading to more descriptive SLAs supporting assured quality of service delivery.

Business and Operational Impact

In summary, driven by the 5G technology innovations and enablement of new verticals and their diversity of future applications, there are numerous economical drivers to push and reshape the overall response by the industry in terms of multi-provider services. We foresee a future where business and technology enablers will be carefully aligned, and provide an agile, efficient, and open multi-service multi-provider solution. Enabled by SDN, we anticipate an evolution into a powerful multi-service platform that goes beyond pure connectivity offerings. This multi-provider platform will also accommodate the needs of, as well as integrate the capabilities of, the emerging NFV and softwarization solutions.

While learning from the IT and cloud industry, the telco and NSP industry will find themselves in an even more uncertain position and challenged with many strategic questions. On one hand, the new technology enablers will allow on-demand trading, flexibility in re-negotiating SLAs, elasticity, and dynamic traffic, resource, and service management. However, which resources should you own, and which resources and services should you buy? What is the best contract duration, and the better roadmap for my service and capability offerings? What kind of partnerships should I develop? How will I best adjust my organization and my operational processes to become an excellent player in such a future [14]? Perhaps the biggest challenge to the network infrastructure and services is "bootstrapping" the basic solution enablers when faced with so many multi-stakeholder coordination issues. On the positive side, solution proposals are now getting more mature, and business attention is rising as the need for solving the challenges are becoming clearer. The use cases and 5G verticals mentioned above are good examples for that.

SDN-enabled solutions can help NSPs evolve their networking solutions in manageable steps according to their business developments and roadmap. Evolving the current IP traffic exchange solutions can complement and augment the current BGP-based operations with managed quality and multi-service inter-NSP 5G-ready traffic exchange services and SLA management solutions. When such a managed and assured quality traffic exchange solution is enabled among a set of initial partners, the solution can scale and grow into a full fledged traffic, resource, and service trading and exchange platform. This holistic multi-domain resource slicing solution will unleash the full potential of 5G.

The value-added connectivity and services that are specific to the end customer are handled by appropriate policies at the SDN-enabled service edge nodes. The traffic flows are then steered onto the appropriate infrastructure automated software quality (ASQ) paths and back-office data centers, according to the application requirements. This way, consistent end-to-end traffic and service handling across domains can be achieved and supported by the intelligence of SDN controllers and the SDN-enabled monitoring and service assurance capabilities. The above anticipated multi-domain service and resource orchestration, hierarchical SLA management, and the need for automated mapping between high-level and lower-level SLAs and their specific configurations and monitoring capabilities will become more crucial as the industry evolves [15].

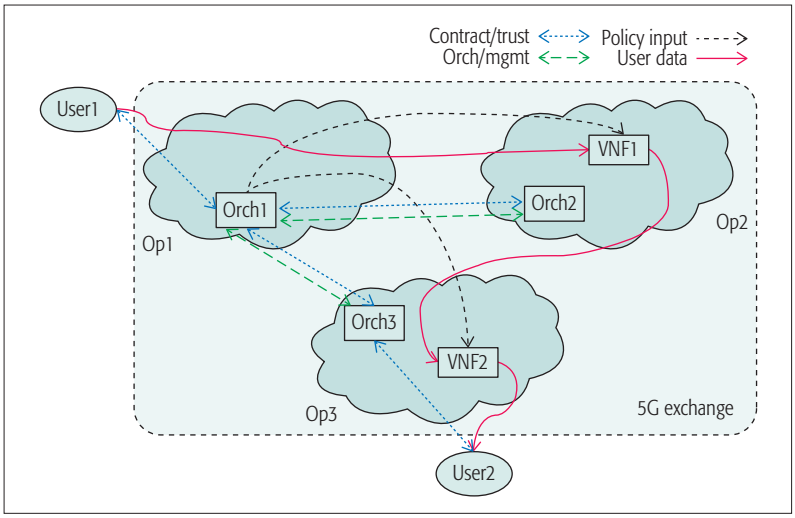


Figure 4. Example: information flows in a 5GEx scenario.

Conclusions

The advent of the 5G era brings with it the promise of value creation by means of a wide range of verticals. On one hand, we have demonstrated how today's best effort Internet is not suitable for the assured quality interconnection these inherently multi-operator services require; other existing solutions are either too domain-specific or have incomplete functionality. On the other hand, we have shown which features of SDN and NFV technologies supplemented with service function chaining serve as key enablers for a novel alternative concept called 5G Exchange (5GEx). We have introduced the 5GEx conceptual architecture and presented how it is able to handle the inter-operator orchestration of composite (network and cloud) resources with the main technical concept of resource slicing. We have also outlined the 5GEx value proposition enabling the creation and trading of complex, high margin services built on top of low margin, commodity building blocks. Furthermore, we have addressed the open question of how to exchange information within the 5GEx framework and provided a roadmap for future research. Finally, we have investigated the business and operational impact of the envisioned solution. We believe that the 5G Exchange is capable of satisfying both the technical and business requirements of future 5G verticals and ushering us into the 5G era.

Acknowledgments

This work has been performed in the framework of the H2020-ICT-2014 project 5GEx (Grant Agreement no. 671636), which is partially funded by the European Commission. G. Biczók and L. Toka have been supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

[1] NGMN Alliance, "5G White Paper," Feb. 2015, https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf, accessed Oct. 27, 2016.
[2] D. Weller and B. Woodcock, "Internet Traffic Exchange: Market Developments and Policy Challenges," OECD Digital Economy Papers No. 207, 2012.
[3] J. Walrand, "Economic Models of Communication Networks," in *Performance Modeling and Engineering*, in Z. Liu *et al.*, Eds., Springer, 2008.
[4] P. Buccirossi *et al.*, "Competition in the Internet Backbone Market," *World Competition*, vol. 28, no. 2, Apr. 2005, pp. 235–54.
[5] EU FP7 Project ETICS, http://www.ict-etics.eu, accessed Oct. 27, 2016.
[6] GSM Assn., "Guidelines for IPX Provider Networks," v. 9.1, May 2013.
[7] OnApp, http://onapp.com, accessed Oct. 27, 2016.
[8] CloudStore, http://govstore.service.gov.uk/cloudstore/, accessed Oct. 27, 2016.
[9] A. Gupta *et al.*, "SDX: A Software Defined Internet Exchange," *Proc. 2014 ACM SIGCOMM*, Chicago, IL, 2014, pp. 551–62.
[10] EU H2020 5GPPP Project 5G Exchange, http://www.5gex.eu, accessed Oct. 27, 2016.
[11] EU FP7 Project UNIFY, http://www.fp7-unify.eu, accessed Oct. 27, 2016.
[12] EU FP7 Project T-NOVA, http://www.t-nova.eu, accessed Oct. 27, 2016.
[13] Ericsson, "5G Systems: Enabling Industry and Society Transformation," Ericsson White Paper, Jan. 2015.
[14] L. M. Contreras *et al.*, "Operational, Organizational and Business Challenges for Network Operators in the Context of SDN and NFV," *Computer Networks*, vol. 92, no. 2, Dec. 2015, pp. 211–17.
[15] P. Wieder *et al.*, Eds., "Service Level Agreements for Cloud Computing," *Springer Science & Business Media*, 2011.

Biographies

Gergely Biczók [S'03, M'11] (biczok@tmit.bme.hu) is an assistant professor at the Budapest University of Technology of Economics, where he received his Ph.D. in computer science in 2010. Previously, he was a postdoctoral researcher at the Norwegian University of Science and Technology, a Fulbright scholar at Northwestern University, and a research fellow at Ericsson Research. His research interests center around the economics of networked systems.

Emmanouil Dramitinos (mdramit@aueb.gr) is a researcher in the Department of Informatics of AUEB. His research interests include network economics and traffic management, mobile networks, auction theory, business modeling, and regulation. He has published high-quality papers in scientific journals and conferences, and is the author of the book *Auction Theory for Telecoms*.

Laszlo Toka (toka@tmit.bme.hu) is currently a researcher with the MTA-BME Information Systems Research Group and an assistant professor at the Budapest University of Technology and Economics. He received his Ph.D. in 2011 at Telecom Paris, and afterward worked at Ericsson Research for three years before rejoining academia. His main research topics lie within software-defined networking, data analytics, and economic modeling of distributed systems.

Poul E. Heegaard [SM'14] (poul.heegaard@item.ntnu.no) received his Ph.D. in telematics from NTNU in 1998. He has been a full professor at NTNU since 2010. His main research interests are performance and dependability modeling and simulations of communication networks, currently focusing on resource optimization and management in distributed autonomous systems in a multi-domain context. He was head of the Department of Telematics (2009–2013), and is now head of the NTNU Quantitative Modelling of Dependability and Performance (QUAM) research lab.

Håkon Lønsethagen (hakon.lonsethagen@telenor.com) is a senior advisor at Telenor Research. He received his B.Sc. in electrical engineering and computer science from the University of Colorado, Boulder in 1987, and an M.Sc. from the Norwegian Institute of Technology in 1988. His current focus is on inter-NSP differentiated and assured services quality and business models, and SDN and NFV. He has participated and been WP and task leader in Eurescom projects, and EU IST DAVID, NOBEL, NOBEL2, ETICS and 5GEx. He has made several contributions to the TM Forum and been an advisory board member of EU research projects SmartenIT and NEAT.

Global State, Local Decisions: Decentralized NFV for ISPs via Enhanced SDN

Alberto Rodriguez-Natal, Vina Ermagan, Ariel Noy, Ajay Sahai, Gideon Kaempfer, Sharon Barkai, Fabio Maino, and Albert Cabellos-Aparicio

Abstract

The network functions virtualization paradigm is rapidly gaining interest among Internet service providers. However, the transition to this paradigm on ISP networks comes with a unique set of challenges: legacy equipment already in place, heterogeneous traffic from multiple clients, and very large scalability requirements. In this article we thoroughly analyze such challenges and discuss NFV design guidelines that address them efficiently. Particularly, we show that a decentralization of NFV control while maintaining global state improves scalability, offers better per-flow decisions and simplifies the implementation of virtual network functions. Building on top of such principles, we propose a partially decentralized NFV architecture enabled via an enhanced software-defined networking infrastructure. We also perform a qualitative analysis of the architecture to identify advantages and challenges. Finally, we determine the bottleneck component, based on the qualitative analysis, which we implement and benchmark in order to assess the feasibility of the architecture.

Introduction

The network functions virtualization (NFV) paradigm enables software-hardware decoupling, flexible deployment of network functions, and dynamic service provisioning [1]. Traditionally, network functions, including firewalls, distributed denial of service (DDoS) filters, TCP optimizers, and so on, are deployed by means of special-purpose hardware appliances. However, the NFV paradigm proposes to virtualize these functions via software in order to dynamically instantiate, move, and destroy them. Complementary to NFV, the software-defined networking (SDN) [2] paradigm has also gained traction in the industry. SDN advocates for decoupling the control plane from the data plane. A central SDN controller centralizes the control and remotely programs the data plane devices. Interestingly, both NFV and SDN serve for network virtualization. While data center and campus networks can use SDN to virtualize network forwarding, Internet service providers (ISPs) can use NFV to virtualize network functions.

ISPs deploy in-network functions to efficiently manage the traffic and offer value-added services to their costumers. In this context, NFV would help to reduce both capital and operational expenses by enabling easier and cheaper deployment and simplified management of network functions. However, bringing NFV to ISP networks presents a unique set of challenges. In addition to performance, manageability, reliability, stability, and security [1, 3], an NFV solution for ISP networks has to consider their large size, the legacy networking hardware already in place, and the heterogeneous traffic generated by ISPs' customers. Therefore, an NFV architecture for ISPs must offer a platform able to scale to a wide range of different workloads and network conditions, while remaining agnostic to the virtual network function (VNF) types required to process the different kinds of traffic.

These requirements dramatically increase the complexity of centralized control. Therefore, we suggest that typical NFV approaches with centralized control fall short for the ISP scale. Those solutions tend to leverage on SDN approaches that centralize — logically — both the state and the control. We advocate that while the network state should be centralized, the control decisions must be decentralized and made locally.

In this article, we analyze these requirements and propose a set of design guidelines to address them. Based on these principles, we describe a novel decentralized architecture that offloads part of the control to an enhanced SDN infrastructure and federates local state through a global database. This makes it possible to make efficient per-flow local decisions based on global knowledge. Therefore, VNFs and client flows can be elastically accommodated. The enhanced SDN is enabled by collocating NFV modules within the SDN controllers and then pushing the controllers close to the data plane devices they control.

Along with the architecture, we present a qualitative analysis that highlights its advantages and challenges. To assess the feasibility of the architecture, we identify its potential bottleneck and provide a possible implementation that we support with experimental performance results.

Scenario Requirements

In addition to common NFV requirements [1, 3], an NFV solution for ISP networks needs to consider the following.

Legacy Hardware

Usually, ISP networks are long-run deployments where there has been a significant investment in network equipment. Contrary to enterprise

The network functions virtualization paradigm is rapidly gaining interest among Internet service providers. However, the transition to this paradigm on ISP networks comes with a unique set of challenges: legacy equipment already in place, heterogeneous traffic from multiple clients, and very large scalability requirements. The authors analyze these challenges and discuss NFV design guidelines that address them efficiently.

Alberto Rodriguez-Natal and Albert Cabellos-Aparicio are with the Technical University of Catalonia; Vina Ermagan and Fabio Maino are with Cisco Systems; Ariel Noy, Ajay Sahai, Gideon Kaempfer, and Sharon Barkai are with Hewlett Packard Enterprise.

Digital Object Identifier: 10.1109/MCOM.2017.1600175

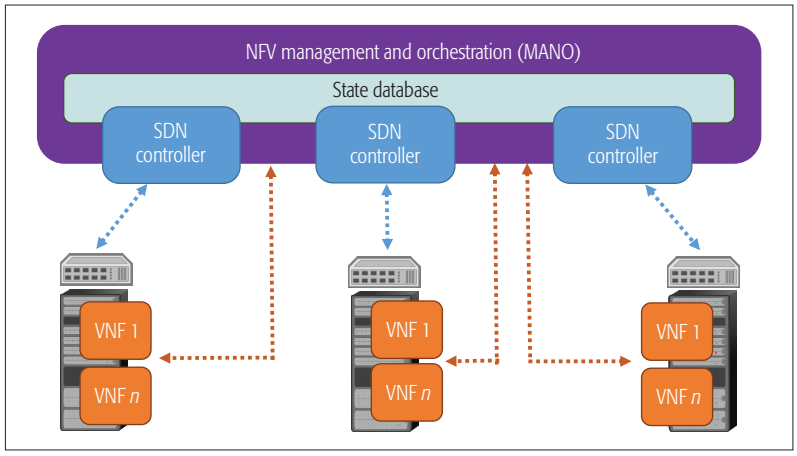


Figure 1. Common centralized NFV approach.

or public infrastructure as a service (IaaS) cloud data centers, which in many cases are greenfield deployments, the networking hardware of ISPs is already in place, and in most cases it is not simple to update, upgrade or replace. The architecture should remain agnostic to the underlying devices and thus be feasible to deploy on top of current networks.

Traffic Heterogeneity

As opposed to the traffic observed in data center networks, the expected traffic in an ISP network will likely come from a wide range of costumers and presents diverse characteristics. Such traffic will require different kinds of processing by large sets of heterogeneous network functions.

Number of Flows

Due to the size of ISP deployments, achieving scalable fine-grained flow processing becomes a major challenge. These networks are typically very large and comprise millions of users. To individually manage large numbers of flows, the architecture needs to scale smoothly.

Global State, Local Decisions

With the advent of SDN and the possibilities of control-data decoupling, network architectures usually centralize the control to ease network deployments. However, we believe that the major challenge of an NFV deployment for ISP networks is that their scale, in terms of both traffic and number of independent subscribers, increases the complexity of keeping logically centralized control scalable.

Existing works on SDN [4, 5] propose a distributed control. We take that approach further, advocating for an NFV architecture where the control is not only distributed but also partially decentralized. We seek to find an optimal middle ground, between the decentralization of legacy networks and the centralization brought by SDN.

We argue that this optimal balance can be achieved by enforcing local control while keeping global state; that is, federating the state generated locally to make the outcome of the local decisions globally available. As a result, the decision on how to act on a flow can be made by the node processing the flow, but taking into account the global state of the system at that time.

For instance, assume a scenario where local controllers locally monitor the load of VNFs. To alleviate the load on a certain VNF, a controller may locally decide to steer some flows to a less loaded VNF. Via a global state database, it knows the load of other instances of that VNF and can locally choose a less loaded one. Then it publishes this decision on the global state database. Thanks to this, if another controller has to locally process those flows, it knows that it has to forward them to the new VNF.

Design Principles

In what follows we propose a set of design guidelines to achieve this partial decentralization as well as to face the other requirements (mentioned earlier) of the ISP scenario.

MANO Disassembling

We refer to the European Telecommunications Standards Institute (ETSI) architectural framework [6] that defines the management and orchestration (MANO) system as the central point for NFV control [7]. The MANO system comprises the global orchestration of the architecture, and the management of the VNFs and the virtualized infrastructure. Albeit this system may be — and in most cases is — physically distributed, it remains a logically centralized point of control, as shown in Fig. 1. We propose to keep the service/functions catalogs and the general NFV orchestration centralized. The management of the virtual resources other than the network (storage, computing) should also remain partially centralized. However, we advocate that the management of the virtual network can be completely offloaded to the infrastructure. Thanks to an enhanced SDN infrastructure, it can be totally decentralized and auto-coordinated. Similarly, the VNF management can also be partially offloaded. The creation and destruction of VNF instances is still coordinated by a centralized entity (e.g., OpenStack.org). However, the monitoring, balancing, and load assignment are decentralized and coordinated directly by the enhanced SDN infrastructure.

SDN Infrastructure Enhancement

The MANO system comprises different instances of an SDN controller to control the network. To achieve an enhanced SDN infrastructure, they must be isolated and pushed close to the data plane devices they control. Furthermore, part of the MANO system itself should be partially distributed over those controller instances to achieve better local control, as shown in Fig. 2. Previous works already discuss collocating NFV and SDN elements with the data plane nodes [8]. We seek to also effectively offload the control to those elements. The state database remains part of the centralized MANO, but it is mostly updated by the decentralized controllers. Controllers collocated with the data plane devices have richer information about the traffic. They can make faster and better decisions than a centralized MANO.

Offload Redundant VNF Functionality

This enhanced SDN infrastructure with decentralized MANO modules offers a general framework where different VNF types can be allocated. Features common among the VNFs (e.g., resilience, load balancing) can be offloaded to the local

MANO modules that use the federated global state to coordinate (Fig. 2). This results in modular, optimized, and compact VNFs, which enables a VNF-agnostic architecture that can efficiently handle the different kinds of traffic expected on ISP networks.

Overlay Encapsulation

An overlay encapsulating traffic over the legacy infrastructure can overcome the constraint of the network equipment already in place. We differentiate three parts of the network, *overlay*, *underlay*, *and outerlay* (as shown in Fig. 3). Overlay is the virtual network instantiated by the architecture through encapsulation. Underlay is the legacy network beneath based on off-the-shelf hardware. Outerlay is the external networks that generate/receive the traffic and connect to the overlay through enhanced SDN edge nodes.

Strong Identity-Location Decoupling

To support the model of a VNF-agnostic overlay-based system with decentralized control, the architecture must enforce a strong decoupling of identity and location semantics and introduce different levels of indirection. We aim to solve NFV challenges by moving the network appliances to the data center and then getting the outerlay traffic there. Identity-location split is required to maintain real-time mappings of VNFs to data center servers, overlay traffic to underlay tunnels, and outerlay clients to offered services.

Architecture

Building on the design principles discussed earlier, we propose the NFV architecture depicted in Fig. 3. To illustrate the architecture, let us assume that an operator has deployed a service to enhance HTTP traffic on its network. This service leverages on using a firewall to check that the subscriber is not accessing a malicious site, and then using a TCP optimizer to boost the transmission performance. In the figure, the VNFs hosted in data center 1 implement firewall functionality, while those hosted in data center 2 are TCP optimizers. In Fig. 3 an HTTP flow from a subscriber arrives at the edge node on the left, which detects that the flow is HTTP traffic subject to enhancement. The edge node queries the global state database to find the VNF chain assigned to that particular flow. Since no chain has been computed, it creates one itself. Based on the current state of the system stored in the database, the edge node decides that the traffic will go through VNF-1 (firewall) and then through VNF-4 (TCP optimizer). This decision is made publicly available by storing the flow-to-VNF-chain affinity in the global state database. Using this global information, the rest of the edge nodes can properly forward the traffic, first across the firewall, then through the TCP optimizer, and finally to the Internet. The rest of this section contains more details on the architecture.

Edge Nodes

Due to the legacy equipment already in place, it is not cost-effective to upgrade all nodes in the network to support the required NFV capabilities. Therefore, the architecture relies on just upgrading the nodes at the network edges (i.e., the ingress/egress points for clients' networks and the locations where the VNFs are hosted). For the VNF case, these edge nodes may be the switches at the top of virtualization racks and/or the gateways of the data centers hosting the VNFs. Given the characteristics of ISP networks, these edge nodes should offer flow granularity for packet processing while keeping line-rate throughput on the data plane and low latency times for the control plane.

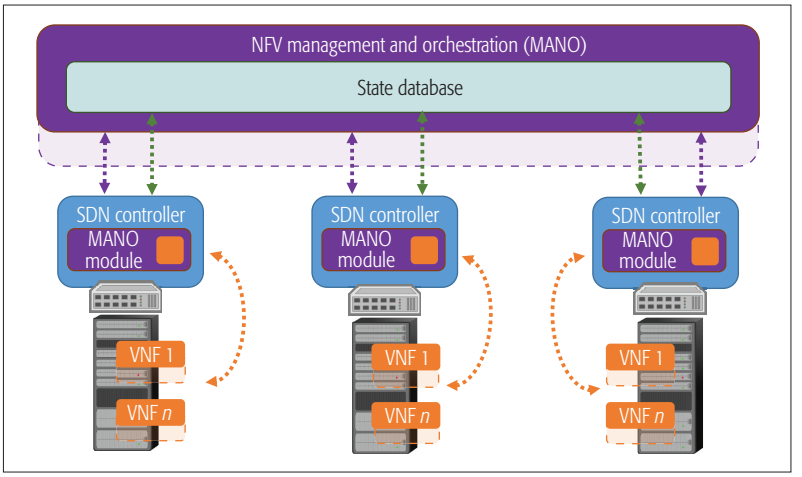


Figure 2. Decentralized NFV based on an enhanced SDN.

To achieve this, we propose the edge node design depicted in Fig. 4. An SDN controller is collocated with a hardware SDN switch to minimize switch-controller latencies. This hardware SDN switch is able to process the traffic at flow granularity and line-rate speed via minimizing the lookup time, that is, only performing exact match lookup over a minimal set of packet fields (e.g., 3-tuple). Any packet that does not hit an exact match entry (i.e., no rule available for its flow) is sent upward to a software SDN switch. Although slower, the software SDN switch allows performing more granular (e.g., 5-tuple) flow lookups and defining as many rules as needed. In general, to classify a packet more fields are needed than to forward it. The software SDN switch contains detailed rules to classify the flow and find the appropriate MANO service module within the SDN controller.

These MANO modules are per-service (e.g., HTTP enhancement) specific software pieces that can assign flows to VNF chains and program accordingly the hardware switch to forward them. Once the software switch hands the flow to the proper MANO module, the module checks if there is already a VNF chain cached that is suitable for the flow. If that is not the case, it uses the controller's database interface to retrieve a suitable chain from the federated information. If no suitable chain exists for that specific flow, the service module has to compute one itself, as described next. After retrieving/computing the VNF chain, the MANO service module uses the controller's southbound interface to program (e.g., via OpenFlow [2]) the exact match rules in the hardware switch. Subsequent packets of the flow will hit the exact match entry and be processed at the hardware level.

Federated Global State

A physically distributed but logically centralized database federates all the state generated locally at the edge nodes (e.g., computed VNF chains).

This database makes the state globally available to the whole infrastructure. It also stores general MANO information (network services catalog, VNF catalog, VNF instances, infrastructure status, etc.) [7]. A summary of the information stored is provided below.

- *VNF class → VNF instances*: The abstract VNF classes that the different service use are instantiated into (and mapped to) specific VNF instances.
- *Flow → VNF chain*: Each flow already processed is mapped to its assigned chain of VNF instances.
- *VNF instance → Instance status*: Per each VNF instance the database stores, its current location, the number of flows assigned to it, and so on.

The database follows a strong location-identity decoupling model to store the information, which allows easy introduction of different levels of indirection. This entitles endpoints to smoothly move across different access networks and allows VNFs to be elastically allocated both inside and outside a data center. For instance, in Fig. 3, VNF 1 (i.e., identity) can be seamlessly migrated from data center 1 to data center 2 (i.e., location) following this schema.

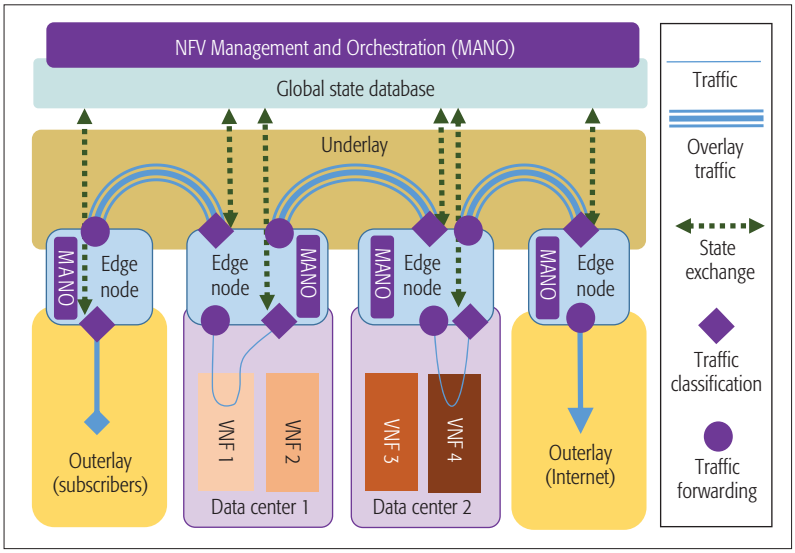


Figure 3. Proposed architecture.

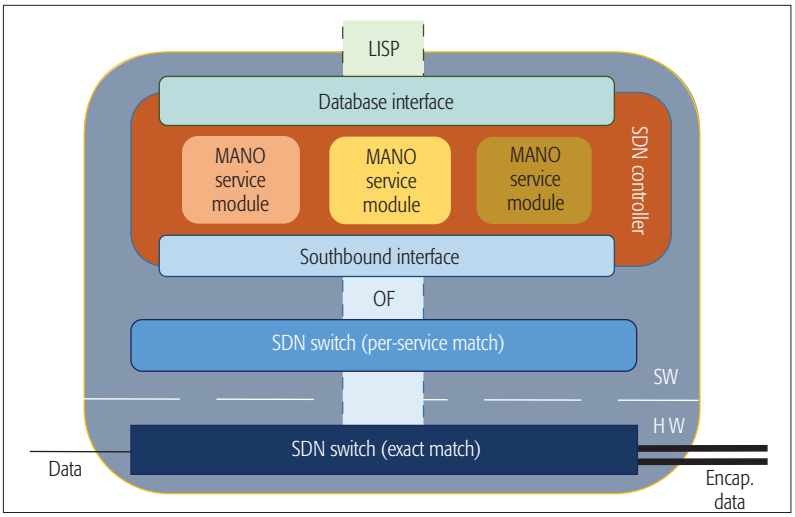


Figure 4. Edge node internals.

For the database implementation, the architecture uses a distributed hash table (DHT) database back-end. Such databases use hashes to index the information and thus offer scalable storage with a delimited query time. In terms of available solutions, Cassandra [9] can fulfill the requirements due to its good availability and excellent scale-out capacity [10]. For the front-end interface to the database, the architecture uses LISP [11, 12], a pull-based protocol that allows retrieving identity-to-location mappings from a central repository. LISP fits the identity-location split model required by the architecture well and offers an interoperable (i.e., IETF-backed) and lightweight mechanism to retrieve state.

Given that the large size of the network leads to considerable state to store, to keep the architecture scalable the state is only pulled on demand by edge nodes. Therefore, in order to report changes and keep the state consistent, the database follows a publish-subscribe mechanism [13]. As an example, if in Fig. 3 VNF 1 has to be moved to data center 2, the edge node at the subscriber's network will be notified and start encapsulating the flow toward data center 2 instead of data center 1.

VNFs

The VNFs are allocated in generic virtualization racks at different data centers with an edge node as ToR switch or data center gateway. Due to the encapsulation and the location-identity split, the VNFs can be dynamically moved across hosts, racks, or data centers. Therefore, the model allows both encapsulating the traffic toward where the VNFs are and/or moving the VNFs close to where the traffic is. All this path computation complexity is offloaded to the MANO service modules at the edge nodes.

In this architecture, the VNFs are unaware of the rest of the system (i.e., they do not know the next hop for a flow). Therefore, the scope of the VNF state is restricted to flow processing. This simplifies the elastic allocation of VNFs and the deployment of new services, since different VNFs from different services can easily be chained.

Ideally, each VNF should perform only one single task, and complex services should be created by chaining different individual VNFs. This enables a flexible system that can scale out in a modular fashion. For instance, if a VNF is experiencing high load, that specific VNF can be scaled out independently without affecting other VNFs in the chain.

In this sense, the architecture trims out redundant logic common to all VNFs and moves it to the distributed MANO modules. Scalability, load balancing, high availability, and so on are decoupled from the VNFs and offloaded to the infrastructure. As an example, a firewall VNF processes packets unaware of any balancing policies. Its local MANO module monitors it and takes care of reassigning flows to properly balance the load among similar firewall VNFs.

Management and Orchestration

The architecture is oriented toward deploying services (e.g., HTTP enhancement) via decentralized MANO modules. The central MANO system

installs these service-specific modules in the edge nodes and programs the software switches to forward the traffic to them. Each service defines the type of traffic to be processed (e.g., HTTP) and the VNF classes to apply (e.g., firewall, TCP optimizer). The central MANO system is in charge of instantiating the VNFs for the service. The decentralized MANO modules build on-demand VNF chains based on available VNF instances, drive the traffic through the VNF chain, and notify the central MANO system when a VNF needs to be migrated.

The service description defines the classes of VNFs to chain, but the service modules decide in real time which is the best VNF chain among all possible VNF instances. For instance, for a real-time analytics service the best chain may be composed of VNF instances placed in low-latency locations, while for an on-the-fly video decoding service the chain may comprise the currently less loaded VNF instances. A computed VNF chain is stored in the database to make it available globally and cached locally to assign it to similar flows in the future. The distributed service-specific MANO modules monitor the traffic and the VNFs, and are synchronized with the federated global state and with the central MANO subsystem. Therefore, they can reassign flows to different chains or recompute chains if required.

Qualitative Analysis

Advantages

An NFV architecture leveraging on SDN comprises several benefits. First, there is flexible and rich control of the network thanks to the SDN controllers. Second, there is inherent support for traffic engineering and load balancing enabled by the SDN fabric. Furthermore, the decentralized NFV architecture that we propose presents a set of additional advantages.

Decentralization Boosts Scale-Out: Since the coordination required among the different parts of the architecture is relaxed, it is easier for these parts to scale out independently. This can be achieved for the architecture as a whole (e.g., adding more edge nodes) or for each component individually (e.g., adding more physical servers to an edge node cluster).

Flow Granularity Even at Large Networks: The optimized flow lookup allows for more flows to be handled per hardware switch and thus reduces the cost of scaling out the edge nodes to allocate more traffic. In general, all architecture components are designed to keep flow granularity despite the network size. Edge nodes process flows in parallel independently, VNFs keep only per-flow state, and the federated database uses a plain namespace with constant access time.

Better Per-Flow Decisions: The decisions on how to process a flow are taken close to the data plane devices carrying the flow itself. Therefore, more and richer per-flow information is available. The flow granularity processing and this detailed per-flow information enable complex per-flow decisions, something that is challenging to accomplish with traditional logically centralized architectures.

VNF Outsourcing: The combination of an architecture that is VNF-agnostic and VNFs that are simple, light, and interoperable enables VNF outsourcing. The VNFs do not need to be specifically developed for the particular NFV system but rather can be developed by third parties and smoothly integrated with other VNFs. The architecture eases the development of such outsourced VNFs, since VNF vendors can leverage on the mechanisms offered by the infrastructure and thus avoid dealing with ISP networks' scalability or availability requirements.

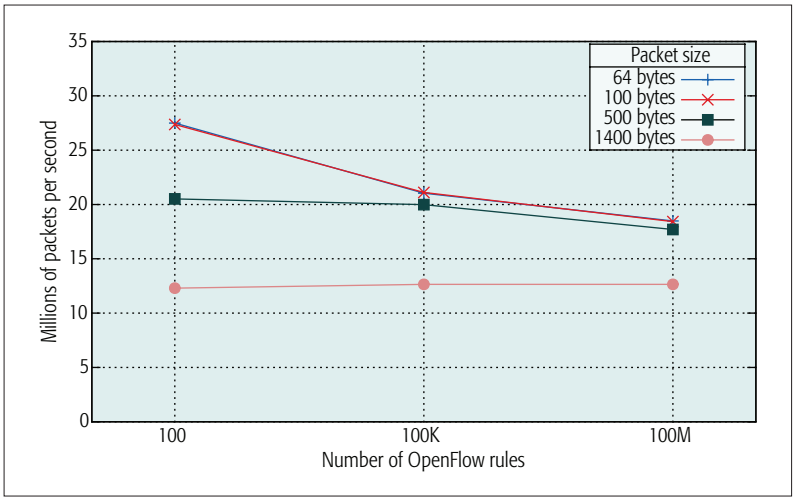


Figure 5. In-house software switch performance with millions of rules.

Challenges

Global State Query Latency and Extra Singling: Relaying on a global state database presents some challenges. First, the state retrieval imposes an inherent latency on the operation of edge nodes. Nevertheless, technologies already available should offer low enough query time. For instance, Cassandra queries take only a few milliseconds even under high loads [10], and an optical underlying transport induces latency on the order of microseconds [14]. Second, the dependency on the global state may result in an overhead of message exchanges. However, caching techniques at the edge nodes and careful design of service requirements may render this extra signaling overhead negligible.

State Inconsistencies on the Decentralized System: The decentralized nature of the architecture may introduce state inconsistencies. However, the state update mechanisms in place make the state eventually converge. Furthermore, local controllers can adapt their control policies to face these temporary state divergences. As an example, the edge node for a particular firewall VNF decides to migrate it to a less loaded location. As a consequence, this new location is published in the database and reported to edge nodes that previously retrieved the firewall's location. While the update propagates through the system, the edge node at the old location will redirect incoming packets addressed to the firewall toward the new location.

Lack of Control for the Underlay Network: The architecture uses the underlay network and has to rely on its correct operation. If that is not the case, the architecture has no control over it and is unable to fix the problem. However, ISP networks will likely have their own troubleshoot-

Contrary to most NFV proposals, the SDN controllers used by the MANO are collocated with their controlees and provisioned with local MANO modules. This enables faster local processing by means of reducing centralization. The architecture proposes a good tradeoff of complexity, performance and scalability by decentralizing some components while keeping a centralized state.

ing and healing mechanisms, as assumed in [5]. Furthermore, in the case of a major connectivity problem, the enhanced SDN infrastructure may be able to transparently detour the traffic thanks to the identity-location split schema enforced.

Edge Node Implementation: The proposed service-based decentralization results in complex edge nodes that need to remain scalable. On one hand, each local MANO module is independent of the others, and its performance is not affected by the number or complexity of other modules. Therefore, scale-out requirements can be met with a cluster-friendly controller (e.g., OpenDaylight.org) able to distribute the load. On the other hand, the hardware switch is agnostic to the service complexity or its number since it only considers independent exact match rules. Thus, it can be scaled out across several hardware devices.

The bottleneck of the system is the software switch. In this case, contrary to the rules allocated in the hardware switch, the rules required to support more services or more complex ones comprise wildcard fields, longest prefix match lookups, and different priorities (since these rules will likely overlap). This makes the complexity of flow classification at the software switch increase nonlinearly with respect to the complexity or number of services. However, the architecture needs a software switch capable of achieving the linear scalability required by the large number of flows expected, despite the flow heterogeneity and the nonlinear complexity faced in the flow classification.

Software Switch Implementation

From the analysis earlier, we conclude that the scalability of the system will be capped out by the performance achieved by the software switch. To cope with the requirements of the ISP scenario, the software switch must be able to keep a high packet throughput despite the number of rules and the heterogeneity of the traffic.

We measured the performance of currently available software switches, particularly Open vSwitch (openvswitch.org), and we were able to achieve 11 million pps using Open vSwitch 2.3.1 Data Plane Developer Kit (DPDK)-optimized (dpdk.org) on a single core with 100 OpenFlow rules and less than 100 traffic flows. This number is similar to the one reported in [15] and, to the best of our knowledge, this is due to caching lookup results for known flows effectively bypassing the OpenFlow lookup tables. When we raised the number of flows to 500,000, a number closer to the ISP scenario, the performance dropped to 300,000 pps. We also observed nonlinear scaling since an 8-core configuration only achieved 1.2 million pps. The hardware used for these tests was similar to the one described later in this section.

To achieve ISP performance requirements we implemented our own in-house software switch, written in C and leveraging on DPDK. It is based on a multithreaded design where all threads have access to the rules from a common memory space. Each thread handles a subset of the flows distributed to it based on 5-tuple hashing performed by the network interface cards (NICs), using receive side scaling (RSS) technology with a queue per thread. The OpenFlow tables are presented as static tables, and updates are performed on a shadow copy of those tables. Periodically, the shadow copy is switched with the active table set and updated with the changes made on the shadow copy, and from there updates commence on the new shadow. The key to the high performance implementation is that for every packet, the relevant rules are fetched into cache memory just in time for lookup. By pipelining the rule prefetching (i.e. handling a few packets in parallel by each thread), the throughput is achieved by always effectively referencing rules that reside in the CPU cache (and not in off-chip memory). As a result, the performance is almost independent of the number of rules.

To measure the scalability of the proposed software switch we performed the following benchmark. We ran the switch on an Intel-based server with dual Intel Xeon E5-2690 2.9GHz 8-core per socket CPU (i.e. 16 total cores) with 128 GB of RAM and a set of 16 interfaces of 10 Gb/s each. We populated the switch with rules ranged from 100 to 100 million, and we generated traffic evenly distributed across all rules (i.e., the traffic was forged to hit all rules at the same rate). Figure 5 shows the packets per second processed by the switch for different numbers of rules and packet sizes. In all cases the delay per packet was constant and around 50 µs. The figure shows how the switch scales almost linearly and achieves the requirements of the architecture.

Conclusion

The architecture presented in this article addresses the challenges of NFV for ISP networks via partially decentralizing the MANO system. Contrary to most NFV proposals, the SDN controllers used by the MANO are collocated with their controlees and provisioned with local MANO modules. This enables faster local processing by means of reducing centralization. The architecture proposes a good trade-off of complexity, performance, and scalability by decentralizing some components while keeping a centralized state.

Acknowledgments

This work has been partially supported by a Cisco research grant, by the Spanish Ministry of Education under grant FPU2012/01137, by the Spanish Ministry of Economy and Competitiveness under grant TEC2014-59583-C2-2-R, and by the Catalan Government under grant 2014SGR-1427.

References

[1] B. Han *et al.*, "Network Function Virtualization: Challenges and Opportunities for Innovations," *IEEE Commun. Mag.*, vol. 53, no. 2, Feb. 2015, pp. 90–97.

[2] D. Kreutz *et al.*, "Software-Defined Networking: A Comprehensive Survey," *Proc. IEEE*, vol. 103, no. 1, Jan. 2015, pp. 14–76.

[3] H. Hawilo *et al.*, "NFV: State of the Art, Challenges, and Implementation in Next Generation Mobile Networks (vEPC)," *IEEE Network*, vol. 28, no. 6, Nov.–Dec. 2014, pp. 18–26.

[4] P. Berde *et al.*, "ONOS: Towards an Open, Distributed SDN OS," *Proc. 3rd Wksp. Hot Topics in Software Defined Networking*, 2014, pp. 1–6.

[5] S. Jain *et al.*, "B4: Experience with A Globally-Deployed Software Defined WAN," *Proc. ACM SIGCOMM 2013*, 2013, pp. 3–14.

[6] "Network Functions Virtualization (NFV); Architectural Framework," ETSI Group Spec. GS NFV 002 (v. 1.2.1), Dec. 2014.

[7] "Network Functions Virtualisation (NFV); Management and Orchestration," ETSI Group Spec. GS NFV-MAN 001 (v. 1.1.1), Dec. 2014.

[8] J. Matias *et al.*, "Toward an SDN-Enabled NFV Architecture," *IEEE Commun. Mag.*, vol. 53, no. 4, Apr. 2015, pp. 187–93.
[9] A. Lakshman and P. Malik. "Cassandra: A Decentralized Structured Storage System," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 2, Apr. 2010, pp. 35–40.
[10] T. Rabl, *et al.* "Solving Big Data Challenges for Enterprise Application Performance Management," *Proc. VLDB Endowment*, vol. 5, no. 12, Aug. 2012, pp. 1724–35.
[11] D. Farinacci, *et al.* "The Locator/ID Separation Protocol (LISP)," IETF RFC 6830, Jan. 2013.
[12] A. Rodriguez-Natal *et al.*, "LISP: A Southbound SDN protocol?," *IEEE Commun. Mag.*, vol. 53, no. 7, July 2015, pp. 201–07.
[13] S. Barkai, *et al.* "LISP Based FlowMapping for Scaling NFV," draft-barkai-lisp-nfv-08, IETF Internet-Draft, June 2016, work in progress.
[14] C. Kachris and I. Tomkos, "A Survey on Optical Interconnects for Data Centers," *IEEE Commun. Surveys & Tutorials*, vol. 14, no. 4, 4th qtr. 2012, pp. 1021–36.
[15] P. Emmerich, *et al.*, "Performance Characteristics of Virtual Switching," *IEEE 3rd Int'l. Conf. Cloud Networking*, Luxembourg, 2014, pp. 120–25.

Biographies

Alberto Rodriguez-Natal (arnatal@ac.upc.edu) received a B.Sc. (2010) in computer science from the University of Leon, Spain, and his M.Sc. (2012) and Ph.D. (2016) from the Technical University of Catalonia, Spain. He has also been a visiting researcher (2014) at the National Institute of Informatics, Japan. He joined Cisco in 2016, where he continues his research on new network architectures with special focus on software-defined networking and programmable overlay networks.

Vina Ermagan (vermagan@cisco.com) is a senior technical leader in the Chief Technology and Architecture Office at Cisco Systems. She joined Cisco in 2008, and has been working on research, design, and development of network virtualization and SDN technologies ever since. She has initiated projects to implement LISP in Open vSwitch, OpenDaylight, and FD.io . She received her M.Sc. in computer science from the University of California San Diego in 2008, and her B.Sc. in computer engineering from Sharif University of Technology.

Ariel Noy (ariel.noy@hpe.com) is responsible for setting the HPE, CSB ConteXtream technical direction, architecture, and ongoing innovation and proofs of concept. He co-founded ConteXtream and served as CTO for nine years. Prior to ConteXtream he was a technical leader at Cisco Systems. He joined Cisco when it acquired Sheer Networks, where he was a co-founder and VP Engineering. He has 11 patents and has a B.Sc. in math and computer science from the University of Tel-Aviv (1996).

Ajay Sahai (ajay.sahai@hpe.com) currently leads Open NFV partner solutions at Hewlett Packard Enterprise. Previously he was director of Technical Marketing at ConteXtream, an SDN startup that was acquired by HPE in 2015. In his 20+ year career he has focused on the wireless/networking/telecom arena and worked in both business and technical roles. He has a B.S.E.E. and an M.B.A. He has a couple of issued patents and several applications pending.

Gideon Kaempfer (gidi@hpe.com) is a Distinguished Technologist and Chief Architect of HPE Contextream, where he has been innovating in the fields of SDN and NFV since 2008. He is a serial entrepreneur, bringing with him 25 years of experience in telecom networking, routing, hardware and software system design, and development. Before joining HPE, he held senior architect or CTO positions at Contextream, Xring Technologies, Axxana, Expand Networks, FlowInspect, SilverKite, and Charlottes Web Networks. He is a TSC member of the OpenDaylight open source SDN platform. He holds an M.Sc. in computer science from the Technion — Israeli Institute of Technology.

Sharon Barkai (sharon.barkai@hpe.com) is a Distinguished Technologist at Hewlett Packard where he specializes in network-databases and functional grids. He holds 13 patents and founded grid startups, Sheer semantic networks (CSCO), Xeround sound netDB, ConteXtream map-assisted overlays (HP), and Fermi Cloud lambda-edge. He received his B.A. in computer science mathematics from Hebrew University, and his M.Sc. in computer science from Columbia University School of Engineering and applied science.

Fabio Maino (fmaino@cisco.com) is a Distinguished Engineer at Cisco Systems, in the Chief of Technology and Architecture Office, where he leads a team that focuses on driving innovation on network virtualization and SDN. He has about 50 patents issued or filed with the U.S. PTO, and has contributed to various standardization bodies including IEEE, IETF, and INCITS. He has a Ph.D. in computer and network security and an M.S. ("Laurea") in electronic engineering from Politecnico di Torino, Italy.

Albert Cabellos (acabello@ac.upc.edu) is an associate professor at Universitat Politécnica de Cataulnya (Department of Computer Architecture). He has been visiting professor at Cisco Systems (San Jose, California), KTH (Kista, Stockholm), Agilent Technologies (Edinburgh, United Kingdom), Massachusetts Institute of Technology (Cambridge), and the University of California Berkeley. He has participated in several research projects funded by companies (Cisco, Intel, Samsung, etc.) as well as publicly funded (FP7, H2020, NSF, and national funding agencies). He is also the co-founder of the Open Overlay Router (http://openoverlayrouter.org) open source project.

Resilient Integration of Distributed High-Performance Zones into the BelWue Network Using OpenFlow

Michael Menth, Mark Schmidt, Daniel Reutter, Robert Finze, Sebastian Neuner, and Tim Kleefass

This work presents the SDN-NeIF architecture, a resilient integration of the HPZs into the NeIF and the legacy infrastructure of BelWue and its connected universities, leveraging OpenFlow and BGP. The concept is validated by a prototype, results from a field trial are provided, and additional benefits of using software-defined networking are discussed.

Abstract

BelWue is the Internet service provider for higher education and research institutions in Baden-Wuerttemberg, Germany. Recently, high-performance zones (HPZs) have been established on major university campuses and interconnected with a high-speed network for innovation and research (NeIF). This work presents the SDN-NeIF architecture, a resilient integration of the HPZs into the NeIF and the legacy infrastructure of BelWue and its connected universities, leveraging OpenFlow and BGP. The concept is validated by a prototype, results from a field trial are provided, and additional benefits of using SDN are discussed.

Introduction

The Internet service provider (ISP) BelWue interconnects 45 higher education and research institutions in Baden-Wuerttemberg, Germany, including 9 university campuses, and altogether about 150 locations. There is a trend toward service centralization among the universities in Baden-Wuerttemberg, that is, some data- or computation-intensive services are offered only by single institutions and are available to others only via the BelWue network. Therefore, a network for innovation and research (NeIF, Netzwerk fuer Innovation und Forschung) has recently been set up to interconnect the university campuses through a flexible optical network with 100 Gb/s wavelengths between neighboring sites, divided into 10 × 10 Gb/s bandwidths. It enables 10 Gb/s point-to-point connections to provide low-latency services between any two campuses, but their overall number is limited by the switching matrices at each site. Therefore, a full optical mesh among all campuses is not feasible. As the existing campus infrastructure of universities is not ready to support these high data rates, and the optical network of BelWue is already on its way to be upgraded to multiple 100 Gb/s, so-called high-performance zones (HPZs) with high-performance hosts (HPHs) are established and directly connected to the NeIF via OpenFlow-capable Ethernet-based border switches (BSs).

The NeIF is already partly used to provide point-to-point connections for special demands to carry data-intensive traffic between university locations. In contrast, the HPZs are currently operated in isolation from the existing (legacy) infrastructure and can be interconnected through point-to-point connections provided by the NeIF (Fig. 1). In this context, the project bwNET100G+ was set up among research groups and the computation centers at the Karlsruhe Institute of Technology, the University of Tuebingen, and the University of Ulm, under BelWue's coordination. Its objective is to make networking within and among universities more flexible, leveraging novel networking technologies like OpenFlow, and to improve transport layer and security aspects to enable university users to benefit from the increased bandwidths in the BelWue network.

The work presented in this article is an outcome of the bwNET100G+ project. It suggests the software-defined networking (SDN)-NeIF architecture, a resilient integration of the HPZs into the NeIF and the legacy infrastructure of BelWue and its connected universities. It uses OpenFlow technology and the Border Gateway Protocol (BGP) for that purpose, and does not require additional hardware, in particular no IP routers for the interconnection of different IP domains. The use of SDN is attractive because it makes networking more flexible and allows for improved security, traffic engineering, and more cost efficiency. This work can be seen as one part of the pre-planing for an evolution toward an SDN-enabled next-generation ISP platform of BelWue.

The rest of the article is structured as follows. We discuss similar activities connecting special zones within and among universities with high-speed networks. We present the conceptual integration of the HPZs into the NeIF and the existing infrastructure of BelWue and the universities leveraging OpenFlow and BGP. We describe the implementation of a prototype and report results from a field trial. We discuss opportunities of the SDN-based HPZ integration. Finally, we conclude this work.

Related Work

We briefly review similar projects that facilitate high-speed communication for scientific applications within or between university campuses. Some of them leverage SDN technology.

McCahill [1] proposed, at an Internet2 technical meeting, a high-speed bypass around a university's core network for special traffic (e.g.,

This work has been supported in the bwNET100G+ project by the Ministry of Science, Research and the Arts Baden-Wuerttemberg (MWK). The authors alone are responsible for the content of this article.

Digital Object Identifier:
10.1109/MCOM.2017.1600177

Michael Menth, Mark Schmidt, Daniel Reutter, and Robert Finze are with the University of Tuebingen; Sebastian Neuner and Tim Kleefass are with BelWue.

scientific data). Departments are connected to the core network with SDN-capable switches. They are interconnected through dedicated high-speed links and decide which traffic to bypass over the high-speed network. The goal is to provide high bandwidth between different departments or different locations of the same university for scientifc data and to relieve the campus core network.

The ESnet Science DMZ [2, 3] defines high-speed zones within universities that are separate from the campus network. Within a university, there may be multiple DMZs with high-performance equipment (e.g., computation and storage servers) and with special security policies and enforcement (e.g., for different projects). The DMZs are used for high-performance scientific applications. They are connected via the university's border router to the campus network and the Internet. SDN technology is used for communication within and among the DMZs of a single university. High-speed connections at 100 Gb/s among the DMZs of a single university enable the use of the resources in different DMZs outside the campus network so that large data volumes do not need to be relayed through the campus network. There is no special high-speed network directly interconnecting the science DMZs of different universities.

SciPass [4] describes a security-enhanced science DMZ. It uses an intrusion detection system (IDS) to classify traffic as trusted and untrusted. Trusted traffic is, for example, scientific data that is exchanged between different universities. SciPass uses OpenFlow switches as load balancers for IDS. After a flow is classified as trusted, the network is configured so that this flow can bypass firewalls and is routed around potential bottlenecks. The goal of this approach is to facilitate the utilization of 100 Gb/s inter-campus connectivity.

Internet2 [5] is a network connecting U.S. education and research centers. It is based on an optical 100 Gb/s backbone and reaches from the U.S West Coast to the U.S. East Coast. The typical uplink of a location is 10 Gb/s. The network can be used, for example, to interconnect science DMZs at different locations. The more general goal of Internet2 is to provide a high-speed network for collaborative applications, distributed research experiments, as well as for grid-based data computation and analytics.

Resilient SDN-Based Integration of HPZs

As illustrated in Fig. 1, the HPZs are connected to BSs that are interconnected through a flexible optical platform — the NeIF. So far, the HPZs are operated in isolation, and the NeIF is only used for point-to-point connections, but not for flexible interconnection of all HPZs. There is no connection to the production environment and the Internet.

In the following, we discuss requirements for the integration of HPZs into the NeIF and the legacy infrastructure. The integration is achieved through the BSs, which are equipped with appropriate forwarding rules. BGP is used to make the HPZs reachable from the campus networks and the Internet. Resilience mechanisms ensure that traffic is rerouted via the BelWue core network if the NeIF fails. Finally, we discuss required information exchange between BelWue and universities.

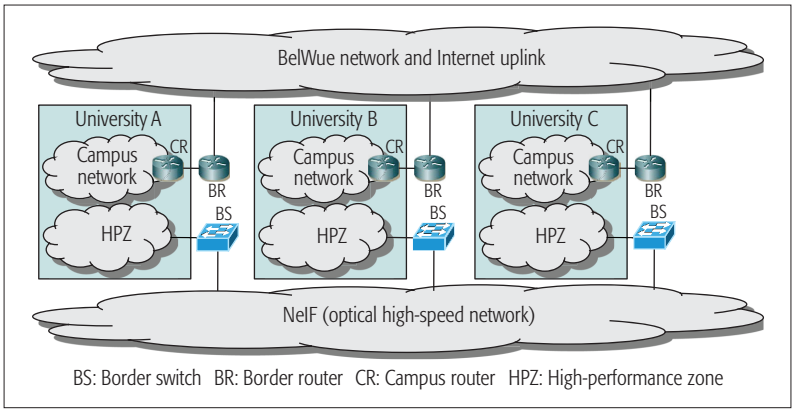


Figure 1. The HPZs belong to the universities. BelWue interconnects them through a flexible, optical network (NeIF). The HPZs are currently operated in isolation from the production infrastructure.

Requirements

As the HPZs belong to different institutions, their equipment belongs to different IP number spaces, so the entirety of all HPZs cannot be operated as a single layer 2 network. Furthermore, the BSs of the HPZs must be managed only by BelWue, while the equipment within the HPZs is managed by the universities. The HPZs must be reachable from university campuses and the Internet, and the reachability information must be communicated in an automated way. The interconnection of the HPZs should require only a few optical links because the ability to establish optical point-to-point links in the NeIF should be retained for special applications. Therefore, only BSs between neighboring sites are connected through single-hop optical links, and BSs relay data among HPZs using packet switching. While university campuses are redundantly connected to the existing BelWue network via two border routers (BRs) and disjoint paths, which are omitted in Fig. 1, the NeIF currently exhibits a tree structure. Therefore, the resilience of the communication among HPZs against link failures should also benefit from their integration into the legacy BelWue network.

Interconnection of Border Switches

A /22 IPv4 and an IPv6 address space are reserved for the entirety of all HPZs, out of which each HPZ receives its own /24 prefix. As illustrated in Fig. 2, any HPZ is connected to the NeIF platform through an OpenFlow-capable BS, which is operated by BelWue and controlled by a BS controller (BSC) run by BelWue. The BS has 10 Gb/s interfaces and a direct link to the campus router (CR), which is typically located in the university's data center. The BS should have a dedicated connection to the BR so that the BR and BS can exchange traffic without forwarding it through the campus network. The BS either directly connects devices within an HPZ or may talk to gateways that hide from the BS the remaining network structure within the HPZ. The BSs have an optical link toward the neighboring BSs in the NeIF. If a single 10 Gb/s channel does not suffice to carry the traffic between neighboring BSs in the NeIF, several 10 Gb/s channels may be bonded using the Link Aggregation Control Protocol (LACP).

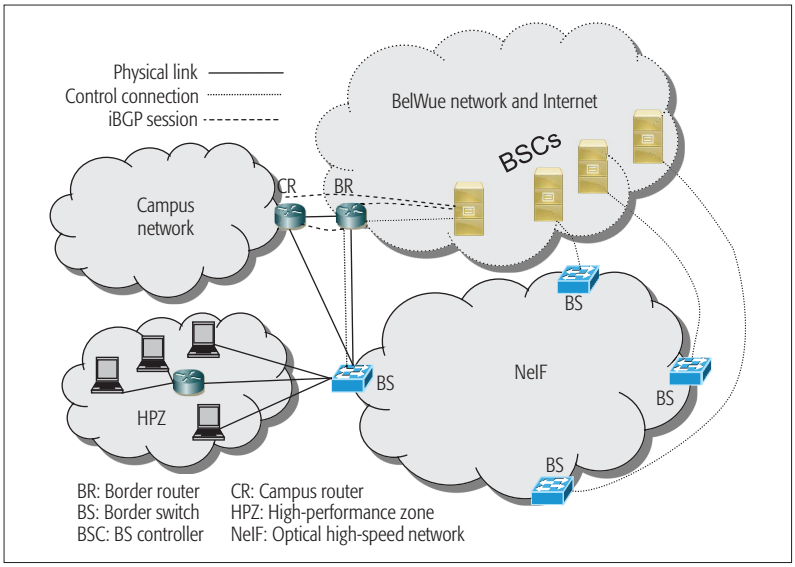


Figure 2. An OpenFlow-capable border switch attaches an HPZ to the campus network, the legacy BelWue infrastructure, and other HPZs over a network of OpenFlow switches.

Forwarding Rules for the BS

The BS requires one static rule to reach its BSC via the BR. In addition, the BS needs rules to act as a gateway for the HPZ. For directly connected devices, dynamically generated forwarding rules are used. The addition of these rules works as follows. A match rule for the /24 HPZ prefix is configured on the BS. It has low priority and triggers the export of the packet header to the BSC. The BSC sends an encapsulated Address Resolution Protocol (ARP) request to the BS, which then broadcasts this request to its neighbors and returns the result to the BSC. Then, the BSC installs a new forwarding rule with higher priority for the requested device on the BS. To avoid ARP requests being broadcast into other HPZs, an additional rule on the BS blocks all non-encapsulated ARP) requests received from the NeIF. If a gateway within the HPZ is connected to the BS, the BS requires a rule to forward traffic toward the prefix behind the gateway. This rule can be configured either statically or with the help of a routing protocol.

The BSC also installs appropriate forwarding rules for the IP prefixes of the other HPZs on the BS. Gateways and directly connected devices have the BS configured as the default gateway, for example, by Dynamic Host Configuration Protocol (DHCP). Moreover, the CR and the BR see the BS as a potential next IP hop. However, if the BS forwards traffic to a neighboring BS in the NeIF, it does not change the source and destination medium access control (MAC) address because packets are forwarded by the BSs only by their destination IP addresses. When forwarding traffic into the HPZ, to the CR, or to the BR, the BS sets appropriate MAC addresses.

HPZ Reachability via BGP

The reachability of the HPZ from the campus and the Internet is achieved through BGP. As the BS is only a switch, it cannot speak BGP itself and requires that the BSC act as proxy to maintain BGP connections with the BR and the CR.[1] The BSC announces the reachability of the /22 prefix of the entirety of all HPZs and the /24 prefix of the local HPZ to the BR and CR with the BS being the next hop. The BR propagates this information further to the Internet and in particular to other BRs. The CR announces itself as the next hop for the university campus to the BSC and BR. Therefore, the BSC installs a rule on the BS to forward traffic destined to the campus to the CR. The BR announces a default route to the CR and the BSC so that it becomes the next hop for the CR and the BS for traffic to the Internet.

Resilience Mechanisms

For resilience purposes, the university campus is connected via two CRs with identical IP addresses to the BR. The Virtual Router Redundancy Protocol (VRRP) [6] is used between them so that they act as one virtual router, which ensures connectivity even if one of them fails. In a similar way, the university uplink is realized via two BRs with identical IP address over disjoint paths to the BelWue core network, and there is a full mesh interconnection among the two BRs and CRs. The figures omit this complexity for the sake of clarity. Because of this arrangement, any BR, CR, or path toward the BelWue core network may fail without compromising the uplink of a location. As a result, the connection between the BS and its BSC via the BR can be considered as highly reliable.

Within the NeIF, the BSs locally detect if a link fails between them and inform their BSCs about this event. In such a case, a BSC withdraws the /22 HPZ prefix via BGP to the CR and the BR. If the link failure is repaired, the BSC is notified and reannounces the /22 HPZ prefix. In case of a failure, the CR and the BR still forward traffic for the local HPZ to the BS, but they forward traffic for all other HPZs via the BR and the BelWue core network to the BRs of the corresponding HPZs. Traffic toward the local HPZ is rerouted by the other BS detecting the failure via its BR, the BelWue core network, the local BR, and the local BS. If that traffic originates in another HPZ, it loads the uplink of a location that is actually not involved in the communication. Therefore, we currently work on controller-to-controller communication so that the BSC can inform other BSCs whose BSs are no longer reachable through the NeIF to reconfigure their BSs and to withdraw the /22 HPZ prefix to their BRs and CRs. As a result, affected traffic toward a local HPZ will be rerouted at its origin. We further work on controller resilience, which is not yet covered by our current prototype.

Information Exchange between BelWue and Universities

In the presented architecture, BelWue controls the BR, the BSs, the BSCs, the NeIF, and universities. Therefore, some information needs to be exchanged between BelWue and universities with attached HPZs.

Devices in the HPZ that are directly attached to the BS respond to ARP requests from the BS so that the BSC can dynamically configure appropriate forwarding rules on the BS. Possibly, this can be simplified, e.g., by an automatic export of ARP mappings from a DHCP server within the HPZ to the BSC. Then, the BSC can install forwarding rules for all devices directly connected to the BS so that the above described mechanism is not needed to set up dynamic forwarding rules.

[1] This functionality is already supported by many controllers, but they can act as proxy only for a single node. Therefore, our design requires a separate BSC for every BS.

The BS needs to be configured with the prefixes that are reachable through attached gateways in the HPZ. The university may communicate this information to BelWue so that the BSC can install appropriate forwarding rules on the BS. As an alternative, the attached gateways may communicate this information through routing protocols via the BS to the BSC in an automated way.

The presented integration makes only a few assumptions about the connection of campus networks and HPZs to BelWue's infrastructure, which are generally met. The only change to the existing campus network is the addition of one BGP neighbor to the CR. All other changes are restricted to the BelWue infrastructure.

Prototype Implementation

We first implemented the concept for the HPZ integration on Mininet and in a local testbed [7]. Then we implemented a prototype on the target platform. As the BelWue and campus networks are production environments, they must not be used for experiments. Therefore, we use only the NeIF and the BSs as physical components and virtualize most other components of the architecture — BRs, campus hosts (CHs), high-performance hosts (HPHs) in the HPZ, and an Internet host (IH)) — as virtual machines (VMs) on servers. Thereby, the prototype can run on the target platform while being isolated from the production infrastructure. In the following, we explain the mapping of physical and virtual components to experimental hardware and briefly describe the virtualization platform.

Prototype Design

The HPZs are currently equipped with test racks. They contain the access to the NeIF, a management VPN, a management switch, an HP ProCurve 3500 switch with 1 Gb/s interfaces, an HP ProCurve 5406 switch compatible to OpenFlow version 1.3 with 10 Gb/s interfaces, and 5 servers with 10 Gb/s interfaces. We utilize the testbed equipment of the HPZs in Tuebingen and Ulm. Access to this equipment is realized via VPN to the management switch, which has direct links to an additional network interface of the equipment. This has the advantage that management traffic does not influence the experiments in the testbed, and access to the components is possible even in case of an error or misconfiguration.

The prototype is illustrated in Fig. 3, which omits the management network. It represents three different sites, each consisting of a virtualized campus network and an HPZ that are interconnected via BSs through the real NeIF. The entities in the figure have site-specific colors. The two HP 5406 switches are subdivided into two and three partitions, respectively. Three partitions are used for the BSs and another two for the Internet connection between the three sites. BS1 is connected via the NeIF to BS2, and BS2 is connected via the NeIF to BS3. This is the central part of the prototype, which runs on physical machines. One HPH per HPZ is implemented as a VM on a dedicated server and connected to the corresponding BS. The BR and CR also run as VMs on a different server and also have a direct link to the BS. Some servers host serveral VMs but have only a single port. Therefore, traffic from the VMs is carried in different VLANs as a trunk

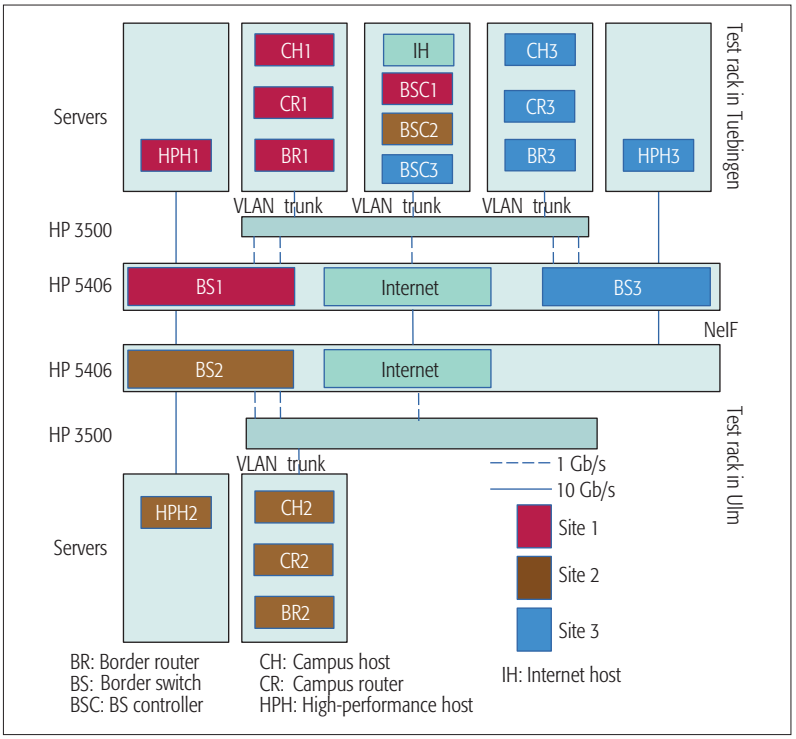


Figure 3. Mapping of physical and virtualized network nodes to experimental hardware.

between the servers and the HP 3500 switch. The HP 3500 switch connects the CR to the CH and the BR, it interconnects the BS with the BR, and it facilitates communication between the BR and the corresponding BSC, as well as among BR1, BR3, and the IH. The virtualized BRs and CRs are based on Quagga, and the BSC leverages the Ryu platform. All IP addresses are statically configured, so a DHCP server is not needed, and debugging is simplified. IP addresses are also directly configured on BSCs, so host discovery is not needed.

Server Virtualization Platform

The servers are equipped with two Intel Xeon E5 processors, 128 GB RAM, three SSDs that are assembled to a RAID5, and two Intel 10 Gb/s network interface cards (NICs), one for management purposes and one for experiments. Each server can host several VMs. As virtualization platform we use KVM as hypervisor in conjunction with qemu. All VMs and hosts run Ubuntu Linux as the operating system with some basic tools for debugging and performance analysis. In [7] we describe how the virtualization techniques are used to build a local testbed for SDN-NeIF. The virtualization concept for the prototype is almost the same. The HP ProCurve 3500 series switches de-/multiplex the VLANs of different VMs of a server to/from different physical switch ports. As a result, the BS and the CR are connected only with untagged VLAN. This is even more realistic because components connect the BSs on dedicated switch ports. Furthermore, the BS is not required to handle VLANs, which keeps the BSC simple.

Field Trial

The prototype in Fig. 3 corresponds to the logical experiment setup in Fig. 4. This field trial involves the real NeIF and and physical BSs, while all other

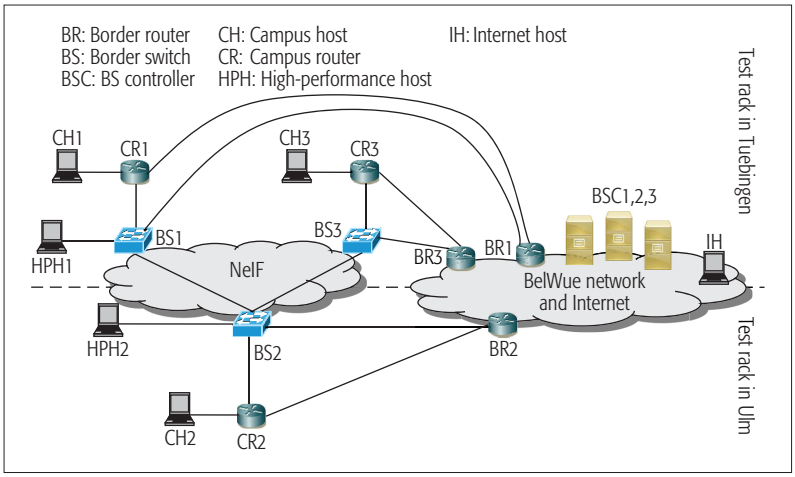


Figure 4. Logical structure of the prototype in Fig. 3.

network entities are virtualized substitutes for production nodes. We performed basic connectivity tests, measured TCP throughput between selected nodes, and checked failover behavior for selected link failures. Additional tests are ongoing.

Basic Connectivity Tests

We used the *ping* and *traceroute* utilities for reachability tests. We verified that:

- CH1 reaches IH via CR1 and BR1.
- CH1 reaches HPH1 via CR1 and BS1.
- CH1 reaches HPH3 via CR1, BS1, the NeIF, and BS3.
- CH1 reaches CH3 via CR1, BR1, BR3, and CR3.
- HPH1 reaches CH1 via BS1 and CR1.
- HPH1 reaches IH via BS1 and BR1.
- HPH1 reaches CH3 via BS1, BR1, BR3, and CR3.
- HPH1 reaches HPH3 via BS1, the NeIF, and BS3.

A local round-trip time between HPH1 and CH1 takes 0.2 ms, a round-trip time over a single NeIF link from HPH1 to HPH2 takes 2.1 ms, and a round-trip time over two NeIF links from HPH1 to HPH3 takes 4.05 ms.

Throughput Tests

We measured throughput from a single TCP connection between selected nodes using *iperf*. HPH1 and HPH3 are connected in the prototype through a path with 10 Gb/s links, and we observed a throughput between them of 9.4 Gb/s. HPH1 and CH1 are connected through a path with a 1 Gb/s bottleneck link, and we measured a traffic rate of up to 960 Mb/s between them. As the traffic between CH1 and CH3 is forwarded via BR1 and BR3, it traverses twice the 1 Gb/s links between the server and the HP3500 switch in the prototype. Therefore, we could only achieve a traffic rate of 480 Mb/s. Thus, we could leverage almost the entire link capacity with only a single TCP connection. Thus, the implementation of the prototype is rather efficient.

Failover Tests

We validated the rerouting for link failures in the NeIF using the ping and traceroute utility. If the link between BS1 and BS2 fails, traffic from HPH2 to HPH1 is rerouted by BS2 via BR2 and BR1, to BS1 and HPH1. Thus, the legacy uplink of the originating location protects the failure. Traffic from HPH3 to HPH1 is first forwarded via BS3 to BS2, which then also reroutes it via BR2 and BR1, to BS1 and HPH1. Hence, the legacy uplink of an intermediate location protects against the failure. As mentioned before, this may be avoided through controller-to-controller communication. If the link failure is repaired, the normal forwarding behavior is restored.

The described recovery process requires that the failure-detecting BS notifies its BSC, which then sends a withdraw to the BR and the CR for the /22 prefix. This approach is only slightly slower compared to the use of a BGP router instead of the BS because the BGP router can immediately withdraw the /22 prefix after failure detection.

Opportunities of SDN-Based HPZ Integration

The proposed integration of HPZs is simple, cost-efficient, and resilient. Apart from that, the SDN-based architecture easily allows campus-to-campus traffic to be carried over the NeIF, facilitates improved security and traffic engineering, and offers perspective on alternative redundant uplinks. Moreover, it can be incrementally deployed, which may be an important step toward a cost-efficient SDN-based wide area network (WAN). We discuss these issues in the following.

Carrying Campus-to-Campus Traffic over the NeIF

The NeIF offers very high transmission capacities that could be leveraged to carry campus-to-campus traffic within the BelWue at a speed of up to 100 Gb/s. This may facilitate the usage of centralized data-intensive services provided by particular university computation centers. Examples are storage and computation clusters. Carrying campus-to-campus traffic requires only reconfiguration of the BS through the BSC to send traffic destined to other university campuses through the NeIF and to announce the other campuses via BGP to the CR. As the number of university campuses within the BelWue is low, configuring the additional required forwarding rules on the BS is feasible.

Improved Security

The use of SDN technology offers flexibility that may be used for improved security in high-speed networks. Currently, we work on OpenFlow-assisted firewall bypassing for selected traffic in high-speed networks and on security concepts for the use of resources in remote HPZs. SciPass [4] leveraged OpenFlow switches to automatically detect scientific data flows that may be safely bypassed around IDS systems. This concept may be reused by bwNET100G+.

Improved Traffic Engineering

With SDN, more flexible traffic management can be supported than with conventional routing. This feature can be leveraged if the CR forwards all outbound traffic to the BS. The BS may be configured by its BSC to forward to the NeIF only traffic from certain applications, or traffic between certain departments or project groups. It is also possible to carry only selected flows (e.g., elephant flows) through the NeIF, while all other flows use the legacy BelWue network via the BR. However, per-flow forwarding may require significantly more forwarding rules on the BS.

Large data transfers between remote HPZs may be automatically scheduled to transmit them at high data rates and to avoid impact on other traffic during busy hours. Such an approach has already been taken by Google [8].

Traffic from the HPZ forwarded via the BR into the legacy BelWue network may overload the existing infrastructure and cause quality of service (QoS) degradation for traffic from the campus network. To avoid that, rate limiting on the link from the BS to the BR may be applied. This feature is available from OpenFlow version 1.3 onward. Such rate limitations may be applied possibly only to traffic from other universities that may emerge in case of failures and rerouting.

Alternative Redundant Uplink

A university campus typically has a redundant uplink. To that end, it is connected with two BRs and two physically disjoint paths to the legacy BelWue network. With the help of the BS and the NeIF, it is possible to provide a redundant uplink with only one BR and the BS at every location. To that end, the CR should forward all traffic to the BS, and the BS forwards desired traffic to the BR. If the BR or the uplink fails, the BSCs need to be notified in some way and reconfigure BRs through iBGP and the BSs such that the affected traffic is carried through the NeIF. Possibly, the traffic can be load-balanced. Requiring only one BR at each location saves operational costs and acquisition costs if a BR has to be replaced.

Cost-Efficient SDN-Based WAN

There is always a run in upgrading to the next magnitude of bandwidth. An IP router usually needs to be replaced as a whole. This is costly in an environment like the BelWue with many sites but only a few institutions per site. Therefore, it is attractive to use the available high bandwidth on the optical layer in a clever way while saving expenses for high-cost devices. An SDN-based WAN is a vision for the future, but is difficult to achieve in practice. ISPs are rather reluctant to introduce SDN technology in their networks because they must ensure stable operation. Their administrators require some time to get familiar with the new control paradigm and develop appropriate debugging tools. The presented approach can be incrementally deployed. First, HPZs may be just interconnected via the NeIF. Then communication with the Internet may be facilitated. Later, resilience may be added. Afterward, the discussed advanced features may be introduced. Incremental deployment offers the possibility to integrate a test network with initially non-critical applications, and use it for production purposes only if sufficient test and operation experience has been gained. Thus, the presented concept simplifies the move toward an SDN-based WAN.

Conclusion

The ISP BelWue interconnects university campuses via a legacy network and their high-performance zones (HPZs) through a high-speed optical network, the NeIF. We present a resilient, OpenFlow-based integration of the HPZs into the NeIF, other existing BelWue infrastructure, and the university campus networks. The proposed SDN-NeIF architecture interconnects different IP domains using OpenFlow switches at a speed of multiple 10 Gb/s. It is designed such that it can be fully controlled by BelWue but gives enough flexibility to cooperating universities. SDN-NeIF is scalable and resilient against failures in the NeIF network with a rerouting speed similar to a pure BGP-based solution. Furthermore, it facilitates improved security and traffic engineering, simplifies a redundant uplink for attached universities, and can be incrementally deployed. The latter is important for the move toward an SDN-based WAN, which is attractive for cost efficiency. The implementation of a prototype and the reported field trial are first steps in that direction.

SDN-NeIF is scalable and resilient against failures in the NeIF network with a rerouting speed similar to a pure BGP-based solution. Furthermore, it facilitates improved security and traffic engineering, simplifies a redundant uplink for attached universities, and can be incrementally deployed.

References

[1] M. McCahill, "SDN, IDM, and Research Computing at Duke," http://meetings.internet2.edu/media/medialibrary/2015/10/19/20151007-McCahill-SDN-IDM-Research-Computing-Duke.pdf, accessed July 20, 2016.
[2] E. Dart *et al.*, "The Science DMZ: A Network Design Pattern for Data-intensive Science," *Proc. Int'l. Conf. High Performance Computing, Networking, Storage and Analysis*, 2013.
[3] ESnet, "Science DMZ," https://fasterdata.es.net/science-dmz/sciencedmz-architecture, accessed July 20, 2016.
[4] GlobalNOC, "SciPass," https://globalnoc.iu.edu/sdn/scipass.html, accessed July 20, 2016.
[5] The Internet2 Community, "Internet2," http://www.internet2.edu, accessed July 20, 2016.
[6] S. Nadas, "RFC5789: Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6," Mar. 2010.
[7] M. Schmidt *et al.*, "Demo: Resilient Integration of Distributed High-Performance Zones into the BelWue Network Using OpenFlow," *Int'l. Teletraffic Congress*, Würzburg, Germany, Sept. 2016.
[8] S. Jain *et al.*, "B4: Experience with a Globally-Deployed Software Defined WAN," *ACM SIGCOMM*, Hong Kong, China, Aug. 2013.

Biographies

Mark Schmidt [S'14] studied computer science at the University of Tuebingen, Germany, and received his diploma degree. Since then, he has been a researcher at the Chair of Communication Networks at the University of Tuebingen , pursuing his Ph.D. His main research interests include software defined networking, OpenFlow, high-speed networks, and virtualization.

Michael Menth [M'05, SM'09] is a professor at the Department of Computer Science at the University of Tuebingen and Chair holder of Communication Networks. His special interests are performance analysis and optimization of communication networks, resilience and routing issues, resource and congestion management, software-defined networking and Internet protocols, industrial networking, and the Internet of Things. He has published more than 150 papers in the field of computer networking.

Daniel Reutter studied computer science at the University of Tuebingen. During his Master's thesis in the the Department of Communications at the University of Tuebingen he worked on the border switch controllers (BSCs) of SDN-NeIF. After obtaining his M.Sc., he joined SySS GmbH and works as an IT security consultant.

Robert Finze obtained a diploma degree in computer science from the University of Tuebingen. Since then he works for the computation center of the University of Tuebingen (ZDV). His main interests are networking architecture and automation.

Sebastian Neuner studied electrical engineering and information technology at the University of Stuttgart, Germany, where he received his B.Sc. and M.Sc. in 2012 and 2015, respectively. He currently works at BelWue, a research and education ISP, as a network architect and engineer, where his main project is upgrading the network to 100G access links and a WSON backbone.

Tim Kleefass earned his diploma in computer science from the University of Stuttgart in 2009. During his studies, he built his first network for the student dormitories with over 2000 access ports. He worked with SWITCH, Zurich, Switzerland, from 2008 to 2009 as part of his diploma thesis and then started working at BelWue, where he planned and operated the BelWue backbone along with all the "interesting stuff" involved. Since 2016 he has been head of network at EXARING AG.

Toward Highly Available and Scalable Software Defined Networks for Service Providers

Dongeun Suh, Seokwon Jang, Sol Han, Sangheon Pack, Myung-Sup Kim, Taehong Kim, and Chang-Gyu Lim

The authors review the state of the art on high availability and scalability issues in SDN and investigate relevant open source activities. In particular, two well-known open source projects, OpenDaylight (ODL) and Open Network Operating System (ONOS), are analyzed in terms of high availability and scalability. They also present experimental results on the flow rule installation/read throughput and the failover time upon a controller failure in ONOS and ODL, and identify open research challenges.

Abstract

Software-defined networking is moving from its initial deployment in small-scale data center networks to large-scale carrier-grade networks. In such environments, high availability and scalability are two of the most prominent issues, and thus extensive work is ongoing. In this article, we first review the state of the art on high availability and scalability issues in SDN and investigate relevant open source activities. In particular, two well-known open source projects, OpenDaylight (ODL) and Open Network Operating System (ONOS), are analyzed in terms of high availability (i.e., network state database replication/synchronization and controller failover mechanisms) and scalability (i.e., network state database partition/distribution and controller assignment mechanisms) issues. We also present experimental results on the flow rule installation/read throughput and the failover time upon a controller failure in ONOS and ODL, and identify open research challenges.

Introduction

Software-defined networking (SDN) is an emerging paradigm that can overcome the limitations in the current network infrastructure. The key idea of SDN is to separate the network control logic from the underlying devices that forward the traffic, and to provide the ability to program the network by means of a logically centralized controller [1]. The centralized SDN controller can easily obtain a global network view, and the performance of a network service can be optimized based on the global network view. Therefore, SDN brings many benefits such as efficient control of network traffic, reduced management cost, and rapid service deployment.

The initial concept of SDN was introduced by the Security Architecture for Enterprise Network (SANE) project [2] of the National Science Foundation (NSF) of the United States in which all routing and access control decisions within enterprise networks are made by a logically centralized server. As a practical instantiation of the SANE project, the Ethane project [3] was introduced and designed a more practical network controller. Based on the success of the Ethane project, a well-known southbound protocol, OpenFlow [4], for communications between the centralized controller and networking devices was devised. In 2011, for more systematic specification, development, and commercialization for OpenFlow, the Open Networking Foundation (ONF) was launched. Until today, various standardization organizations such as the Internet Engineering Task Force (IETF), Internet Research Task Force (IRTF), and International Telecommunication Union Telecommunication Standardization Sector (ITU-T) are working on the standardization related to SDN technologies.

While SDN was mostly targeted at data center networks or campus networks in its initial phase, SDN technologies are evolving toward SDN 2.0, which is targeted at carrier-grade networks or service providers' networks. Given the mission-critical and large-scale nature of carrier-grade networks, the control plane of SDN should be designed in a highly available and scalable manner. In this context, constructing the control plane with a single SDN controller can cause the following problems:

- A single SDN controller can become a single point of failure.
- The size of networks that can be handled by a single SDN controller is limited.

Therefore, more than one SDN controllers should be managed as a cluster, and network services/data provided by a single controller should be replicated across the cluster for high availability (HA). At the same time, for high scalability (HS), workloads should be fairly distributed across the cluster. To address these HA and HS issues, several works have been conducted in the literature, and open source communities are very active in developing highly available and scalable SDN controllers.

In this article, we first review the state of the art on HA/HS issues in SDN and survey relevant open source activities. In particular, two well-known open source projects, OpenDaylight (ODL) and Open Network Operating System (ONOS), are analyzed in terms of the HA/HS issues. We also carried out an experimental study for ONOS and ODL to show the flow rule installation/read throughput depending on the

This work was supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (B0101-15-233: Smart Networking Core Technology Development and B0190-15-2012: SDN/NFV Open-Source Software Core Module/Function Development).

Digital Object Identifier: 10.1109/MCOM.2017.1600170

Dongeun Suh, Seokwon Jang, Sol Han, Sangheon Pack, and Myung-Sup Kim are with Korea University; Taehong Kim was with the Electronics and Telecommunications Research Institute (ETRI), Korea, and is now with Chungbuk National University; Chang-Gyu Lim is with the Electronics and Telecommunications Research Institute.

0163-6804/17/$25.00 © 2017 IEEE

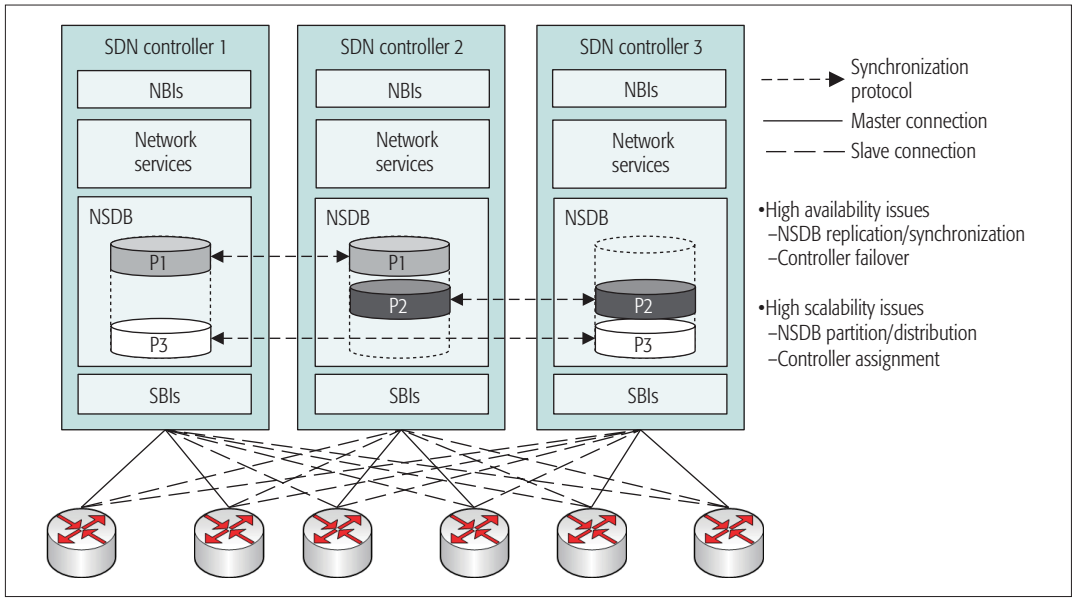


Figure 1. A general clustering architecture with three synchronized SDN controllers.

cluster size and the failover time upon a controller failure.

The remainder of this article is organized as follows. In the next section, key issues for high availability and scalability in SDN are identified, and relevant works are summarized. After that, how to address those issues in ODL and ONOS are discussed, and experimental results on their performance are given. Finally, this article concludes with open challenges.

High Availability and Scalability Issues in SDN

As mentioned before, in order to construct a highly available and scalable control plane, multiple SDN controllers should be managed as a cluster. Figure 1 shows a general clustering architecture with three synchronized SDN controllers. An identical set of network services (e.g., forwarding service, network access control, etc.) are running in the controllers while their network states are stored in the distributed network state database (NSDB).

In Fig. 1, database partition/distribution and replication/synchronization techniques are deployed to the NSDB, which are well-known techniques for high scalability and availability, respectively, in a distributed database. In order to distribute data access load among controllers, the NSDB is logically partitioned into three partitions (i.e., P1, P2, P3), while replicas of partitions are fairly distributed across the cluster. Also, to cope with a controller failure, each partition is replicated into two replicas, and synchronization among the replicas is supported to maintain consistency.

Meanwhile, a master/slave model is leveraged for controller-to-device connections. That is, a device in the data plane establishes multiple connections toward controllers (i.e., master/slave connections) where a controller who has a master connection of the device is only permitted to control the device. Upon a controller failure, one of the slave connections becomes a new master connection. Also, for load balancing, each controller is assigned a subset of master connections of devices.

In such environments, four technical issues on high availability and scalability can be identified:

1. How to partition the NSDB and distribute replicas of NSDB partitions
2. How to replicate NSDB partitions in a consistent manner
3. How to recover master connections from a controller failure
4. How to assign master/slave connections for devices

Note that 1 and 4 are related to HS, while 2 and 3 are related to HA. In the following, we elaborate on each issue and summarize existing works related to the corresponding issue.

Network State Database Partition

By a partitioning strategy, the NSDB is divided into multiple partitions. Each NSDB partition can have multiple replicas for high availability, and replicas are distributed across multiple controllers for high scalability. As partitioning and distribution strategies can affect scalability, it should be carefully designed.

Özsu *et al.* [5] presented three basic partitioning strategies in the relational database: round-robin, hash, and range partitioning. In round-robin partitioning, with n partitions, the ith tuple in insertion order is assigned to partition $k = (i \bmod n)$. Hash partitioning applies a hash function to some attributes that yield the partition number. Range partitioning distributes tuples based on the value intervals of some attributes. Also, Özsu *et al.* [5] introduced a general fragment distribution model, which minimizes the total cost of query processing and storage on each site under the constraints of query response time and storage/query processing capacities of sites.

Krishnamurthy *et al.* [6] investigated the dependency between the application state partition and the devices. They derived the optimal assignment of switches and state partitions to distributed controllers that minimizes inter-controller communications.

By a partitioning strategy, NSDB is divided into multiple partitions. Each NSDB partition can have multiple replicas for high availability, and replicas are distributed across multiple controllers for high scalability. As partitioning and distribution strategies can affect scalability, it should be carefully designed.

In a large-scale network consisting of a number of devices, if only a few controllers act as master, the controllers might be overloaded and result in performance degradation. Therefore, the master controller should be carefully assigned for each device so that the load from the devices can be fairly distributed.

	NSDB partition (granularity/distribution)	NSDB synchronization (protocol/consistency)	Controller failover	Controller assignment
ONOS EventuallyConsistentMap	Store/NA	Anti-entropy protocol/ weaker consistency	Master/slave	Same number of master connections per controller
ONOS ConsistentMap	Map entry/hash-based	Raft protocol/strong consistency		
ODL DistributedDataStore	YANG module/ administrator-defined	Raft protocol/strong consistency	Master/slave	Same number of master connections per controller

Table 1. High availability and scalability approaches in ONOS and ODL.

Network State Database Synchronization

In a database field, synchronization strategies are used to provide consistency between replicas and can be classified into two types:

1. Synchronization strategy with strong consistency, which guarantees all replicas to return the same value when queried with an object
2. Synchronization strategy with eventual consistency, which guarantees that if no new updates are made to the object, eventually all accesses return the last updated value [5]

Meanwhile, strong consistency can only be achieved at the cost of additional latency, and different degrees of consistency can be considered. Thus, the synchronization strategy among replicas should be carefully designed.

Ongaro *et al.* [7] proposed a synchronization strategy with strong consistency, called the Raft consensus algorithm, in which all read/write requests can only be handled by a unique leader replica elected from among candidate replicas, and the read/write requests on any replicas are forwarded to the leader to be processed. For processing of write requests, the agreement among the replicas is mandatory to guarantee strong consistency. Also, in order to ensure that the leader replica is alive, the leader replica periodically sends Raft heart-beat messages to the follower replicas. If one of the follower replicas does not receive any response from the leader replica for a pre-defined election timeout, it requests a new leader election, and the replica with the most votes is elected as a new leader replica.

Botelho *et al.* [8] proposed a novel SDN architecture that focuses on highly available and strongly consistent data storage by using state-of-the-art consistent replication techniques. Botelho *et al.* [9] also developed a fault-tolerant controller architecture with a data store based on a replicated state machine and a lease management algorithm selecting a master controller for fault-tolerant SDNs.

Controller Failover

In OpenFlow 1.2 or higher, multiple controllers for a single device are allowed for reliability, and a device maintains one of the following roles for each controller: equal, slave, and master. A device sends all OpenFlow asynchronous messages to its master controller and accepts OpenFlow controller-to-switch messages from its master controller. On the other hand, a device does not send any asynchronous messages to its slave controller and allows read-only access for it. Similar to the master controller, the equal controller has full access to the device.

The master/slave connection management is responsible for assigning new master controllers for orphan devices (i.e., devices that have lost their connections with their master controllers) while satisfying the constraint of at most one master controller. By providing such a mechanism, the number of dropped asynchronous messages from the orphan devices can be minimized.

Obadia *et al.* [10] proposed two controller failover strategies in which active neighbor controllers take over the control of orphan OpenFlow switches:

1. A Greedy strategy where neighbor controllers take over orphan switches from which they can receive messages
2. A pre-partitioning approach where neighbor controllers proactively exchange information with each other on which switches to take over upon a controller failure

Controller Assignment

Master/slave connection management is responsible for coordination of master connections of devices. In a large-scale network consisting of a number of devices, if only a few controllers act as masters, the controllers might be overloaded, resulting in performance degradation. Therefore, the master controller should be carefully assigned for each device so that the load from the devices can be fairly distributed.

Dixit *et al.* [11] revealed that a static mapping between a network device and a controller can result in lack of dynamic load adaptation capability and proposed a switch migration protocol that can dynamically expand or shrink the controller pool depending on the traffic condition.

Open Source Approach for High Availability and Scalability: ONOS vs. ODL

The development of SDN controllers is led by open source communities such as ONOS and ODL, and high availability and scalability are two important issues in ONOS and ODL. In this section, we briefly introduce ONOS and ODL, and explain how to address the aforementioned issues in ONOS and ODL. Key comparison results are summarized in Table 1.

High Availability and Scalability in ONOS

Figure 2 shows an ONOS clustering architecture that consists of an identical set of network services running in each ONOS instance (only shown for ONOS1 for simplicity) and a middleware component, called Distributed Core, that

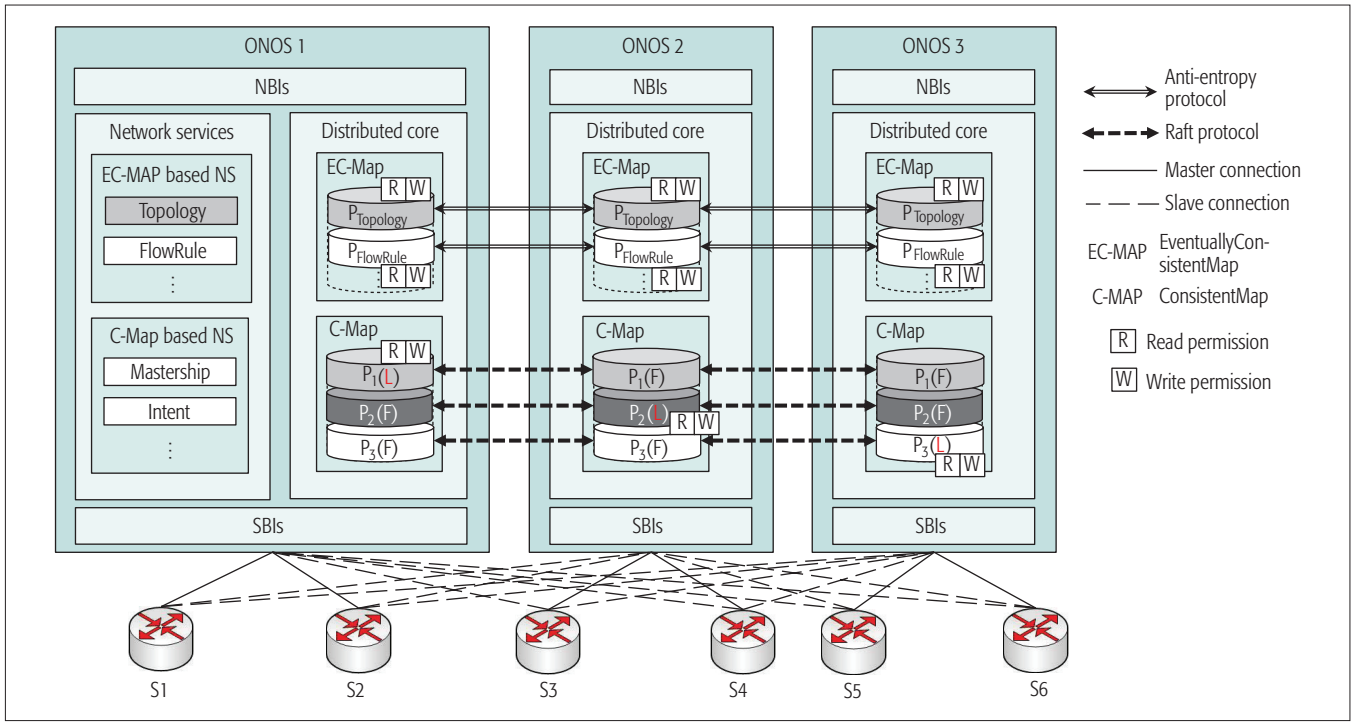


Figure 2. An example of NSDB partitioning/synchronization in ONOS.

manages distributed operations across the cluster and provides two types of NSDBs with different partitioning and synchronization strategies:

1) EventuallyConsistentMap
2) ConsistentMap

Each network service implements its own data storage called Store by means of the two types of NSDBs and accordingly can be classified by which type of NSDB it uses. As shown in Fig. 2, topology and flow rule services are based on EventuallyConsistentMap, while mastership and intent services are based on ConsistentMap. Also, ONOS allows each device to have multiple connections to multiple ONOS controllers, and the master/slave connection management is provided for load balancing and controller failover.

In order to detect a controller failure, ONOS leverages a φ-accrual failure detector [12] where controllers exchange heartbeat messages periodically to keep track of the suspicion level of failure φ for each controller. Each ONOS controller calculates the values of φ for other controllers as $\varphi = -\log10(1 - F(t))$ where $F(t)$ is the cumulative distribution function of a normal distribution with mean and standard deviation estimated from historical heartbeat inter-arrival times t. If a value of φ for an ONOS controller is greater than a pre-defined threshold Φ, the controller is considered as failed.

Network State Database Partition: EventuallyConsistentMap is partitioned into S partitions where S denotes the number of network services, and each partition contains data of each network service (i.e., Store). As shown in Fig. 2, P_{Topology} and P_{FlowRule} contain data of topology and flow rule services, respectively. Meanwhile, all partitions of EventuallyConsistentMap are fully replicated into all controller instances joining the cluster.

On the other hand, ConsistentMap is partitioned into n partitions where n is configurable by the administrator and set to the number of controllers in the cluster by default. Data to be contained in each partition is determined by a hash value of each ConsistentMap entry's key where the hash range is $[1, N]$. For example, in Fig. 2, P_1 contains ConsistentMap entries whose hash values are 1. For ConsistentMap, each partition has R replicas where R is a configurable parameter, and each replica is assigned to the controller that has the least number of replicas. Figure 2 shows a case when $R = 3$.

Network State Database Synchronization: For EventuallyConsistentMap, replicas of each partition are synchronized based on the anti-entropy protocol [14]; it provides weaker consistency guarantee in return for superior read/write performance. All replicas of a partition of EventuallyConsistentMap can handle read/write requests. Specifically, read requests are handled only by the local replica, whereas write requests are handled by the local replica first and the updates are subsequently propagated to other replicas. In order to resolve write conflict and ensure replica convergence, upon receiving an update event for an EventuallyConsistentMap entry, a replica assigns the logical timestamp to the update. Then the update is committed into the EventuallyConsistentMap entry and, in parallel, broadcasted along with the timestamp to other replicas. Upon receiving the broadcasted update event, each replica checks if it has a more recent update for the entry. If the received timestamp is older, it discards the update. Otherwise, the update is committed into its EventuallyConsistentMap entry. By doing so, the system state across all replicas eventually converges to the correct state.

Also, in order to promptly synchronize a replica of a newly joining or restarted controller, at fixed intervals, each replica randomly selects

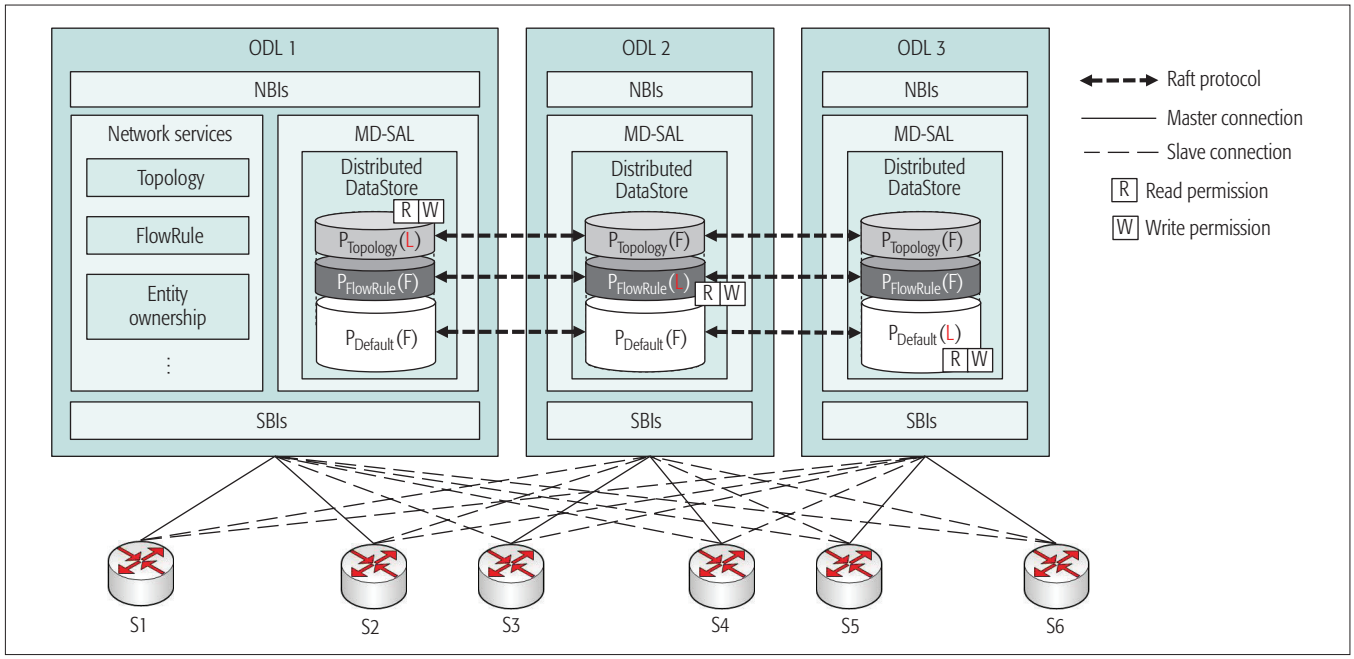


Figure 3. An example of NSDB partitioning/synchronization in ODL.

another replica, and they both synchronize their states. If one replica is aware of more recent EventuallyConsistentMap entries that the another replica does not have, they exchange the entries. For example, in Fig. 2, consider that a topology service running in ONOS3 received a topology update event from S6 and issued a write request to the topology state. The write request is committed into the local replica (i.e., P_{Topology} in ONOS3) immediately and, in parallel, propagated to other replicas. Meanwhile, before the write request is committed into P_{Topology} in ONOS1 and ONOS2, read requests to P_{Topology} in ONOS1 and ONOS2 may observe stale topology state.

For ConsistentMap, replicas of each partition are synchronized based on the Raft protocol [7], which provides strong consistency at the cost of inferior read/write performance. For example, in Fig. 2, consider that an intent service running in ONOS1 has issued a write request to the ConsistentMap entry contained in P_3. Since the local replica is a follower (i.e., P_3(F) in ONOS1), the request cannot be handled locally and should be forwarded to the leader replica (i.e., P_3(L) in ONOS3). Then, after obtaining agreements among the replicas, the leader replica commits the write request.

Controller Failover: Upon an ONOS controller failure, other ONOS controllers in the cluster detect the failure by the failure detector, and re-assign a new master controller for orphan devices. The newly elected master controller for each device sends a role request message to the device to set its role to the device as master, and if successful, it receives a role reply message from the device. As a result, all the orphan devices can recover their master connections.

Controller Assignment: When a new device is connected to multiple ONOS controllers, its master controller is set to the controller that has the smallest number of master connections. By doing so, the number of devices that each controller serves as the master becomes balanced.

High Availability and Scalability in ODL

Figure 3 shows an ODL clustering architecture that consists of an identical set of network services running in each ODL instance (only shown for ODL1 for simplicity) and a middleware component, called the model driven-service abstraction layer (i.e., MD-SAL), that manages distributed operations across the cluster and provides an NSDB called DistributedDataStore. Different from ONOS, each network service in ODL models its data as a form of the YANG module [13] where YANG is a data modeling language. Based on the YANG modules, DistributedDataStore is constructed to store data of network services. Also, similar to ONOS, ODL provides the master/slave connection management and uses a φ-accrual failure detector.

Network State Database Partition: In ODL, the administrator partitions DistributedDataStore into several partitions and selects which YANG module is to be contained in the partitions. There is one special partition called Default Shard, which contains all data except the data defined by the selected YANG modules by the administrator. As shown in Fig. 3, YANG modules of topology and flow rules can be selected by the administrator, and DistributedDataStore can be partitioned into three partitions accordingly. Each partition is replicated into R replicas where R is configurable by the administrator. Figure 3 shows a case when R = 3. Similar to ONOS, each replica is assigned to the controller with the least number of replicas.

Network State Database Synchronization: As in ConsistentMap in ONOS, ODL uses the Raft protocol [7] for synchronization between replicas of a partition. Different from ONOS, all network services in ODL are provided with the Raft protocol for synchronization between their replicas. For example, in Fig. 3, consider that a topology service running in ODL3 received a topology update event from S6 and issued a write request to the topology state. Since the local replica is a follower (i.e., P_{Topology}(F) in ODL3), the request cannot

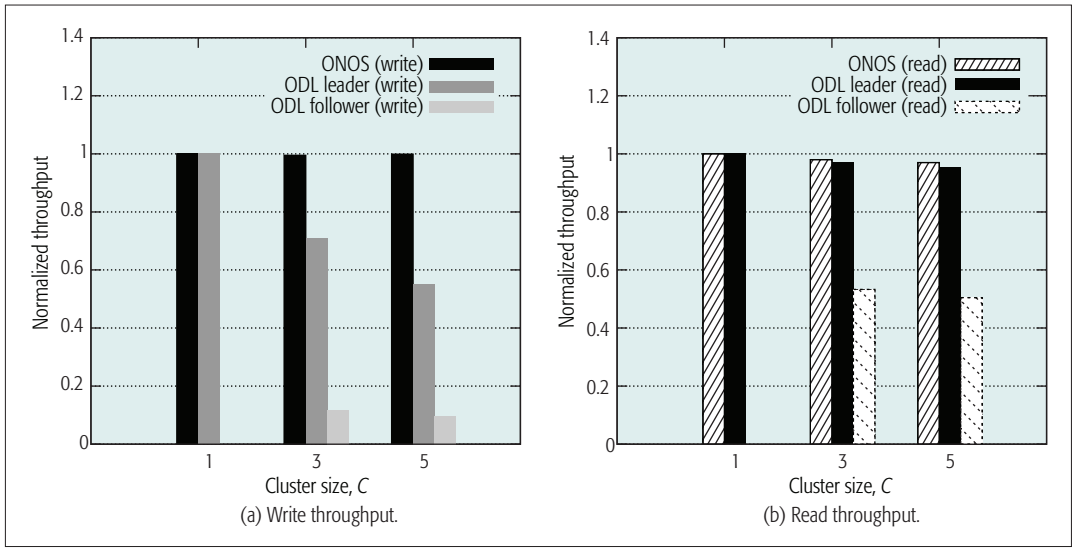


Figure 4. Normalized flow rule installation/read throughput: ONOS vs. ODL.

be handled locally and thus should be forwarded to the leader replica (i.e., P_{Topology}(L) in ODL1); upon obtaining agreements among replicas, the leader replica commits the write request. Therefore, although the consistency between topology state replicas is guaranteed all the time in ODL, read/write performances can be degraded.

Controller Failover and Controller Assignment: The controller failover and controller assignment mechanisms in ODL are similar to those of ONOS; therefore, we have omitted the corresponding descriptions.

Experimental Results

For comparative study, we evaluate the performance of ONOS and ODL in terms of flow rule installation/read throughput and controller failover time. We run each ODL controller in 6 GB RAM and a 2 CPU core virtual machine (VM) with Ubuntu 14.04.2 LTS and each ONOS controller in 6 GB RAM and a 2 CPU core VM with CentOS 6.7, respectively. In terms of version of ODL and ONOS, we use ODL lithium-SR3 distribution and ONOS-1.4 (Emu) distribution. Each experiment is repeatedly carried out to obtain reliable sample values, and the results are obtained by averaging the sample values.

To evaluate the flow rule installation throughput, we run *C* ONOS/ODL controllers, assign one ONOS/ODL controller as a master controller of nine devices, and install 500 flow rules per device through the master controller. Also, the partition that contains the flow rule state is fully replicated into all controllers in ONOS/ODL. For ODL, flow rules are generated, contained in HTTP POST messages, and then transmitted through the northbound REST application programming interface (API) of the master controller to add/delete the bulk of the flow rules into the Inventory Shard that contains flow rules. After that, the OpenFlow controller service in ODL is notified of the data change in Inventory Shard and installs the flow rules to OpenFlow switches. On the other hand, a flow rule installation request tool is used in ONOS.[1] Specifically, the tool requests installation of flow rules to the flow rule throughput test application running in the master controller. After that, the test application creates flow rules randomly and writes the flow rules into the local FlowRule Store, which is based on the EventuallyConsistentMap. As a sequel, the OpenFlow controller service in ONOS installs the flow rules to OpenFlow switches. Since the flow rule installation procedures of ONOS and ODL are different from each other, flow rule installation throughput values of ONOS and ODL are not directly comparable. Therefore, we consider the normalized flow rule installation throughput for ONOS and ODL where the flow rule installation throughput values of ONOS and ODL are normalized by the flow rule installation throughput values when *C* is 1 for ONOS and ODL, respectively.[2]

Figure 4a shows the normalized flow rule installation throughput depending on the number of controllers in the cluster, *C*. For ONOS, it can be seen that the throughput is rarely affected by *C*. This is because since all replicas can handle read/write requests, upon receiving flow rule installation requests, the master controller updates its local replica first. On the contrary, in ODL, two different results are obtained when:

1. The master controller contains a leader replica.
2. The master controller contains a follower replica.

For case 1, agreement among the follower replicas is mandatory before committing the flow rules into the leader replica. Consequently, the latency for committing flow rules increases and the throughput decreases with the increase of *C*. Also, in case 2, degraded throughput is observed as *C* increases due to the increased commitment latency. Moreover, case 2 shows drastically reduced throughput compared to case 1. In case 2, when the master controller receives flow rule installation requests, it forwards the requests to the controller that contains the leader replica. Only after the flow rules are committed into the leader replica, the master controller is notified with the flow rule changes remotely. This forwarding of flow rule installation requests and remote flow rule change notifications cause additional latency; therefore, case 2 shows drastically reduced throughput.

Meanwhile, there is a trade-off between

The newly elected master controller for each device sends a role request message to the device to set its role to the device as master, and if succeeded, it is replied with a role reply message from the device. As a result, all the orphan devices can recover their master connections.

[1] The northbound REST API for the bulk flow rule installation is under development and unavailable in ONOS-1.4 (Emu) distribution.

[2] The flow rule installation throughput values of ONOS and ODL when *C* is 1 are 13,436.7 flows/s and 4672.4 flow/s, respectively.

As ODL provides strong consistency for flow rules, the flow installation throughput can be degraded compared to ONOS which provides eventual consistency for flow rules. However, the eventual consistency for flow rules in ONOS may potentially present a temporal inconsistency and cause undesired behavior of network services that subscribe flow rules.

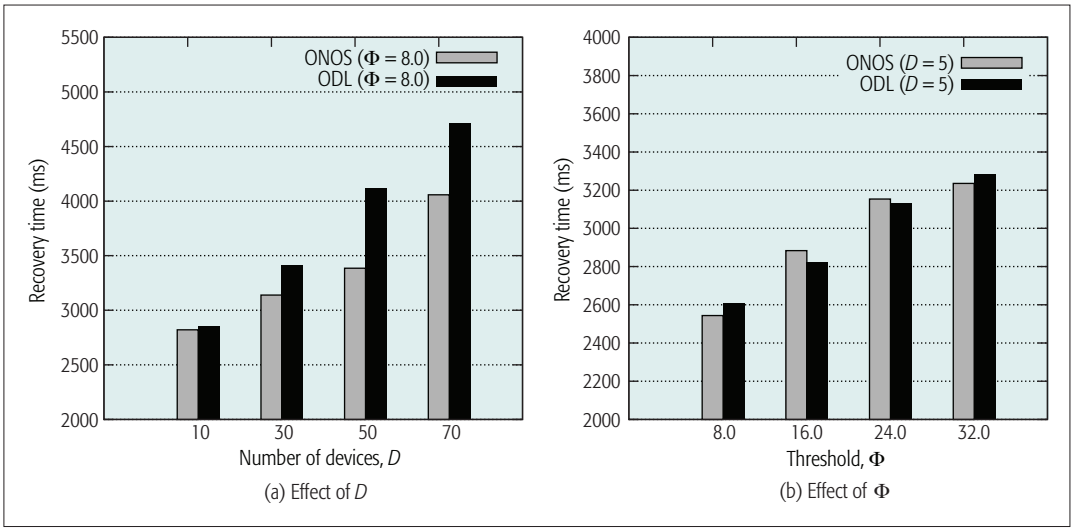


Figure 5. Controller failover time: ONOS vs. ODL.

ONOS and ODL in terms of the flow rule consistency and the flow rule installation throughput. As ODL provides strong consistency for flow rules, the flow installation throughput can be degraded compared to ONOS, which provides eventual consistency for flow rules. However, the eventual consistency for flow rules in ONOS may potentially present a temporal inconsistency and cause undesired behavior of network services that subscribe flow rules.

To evaluate the flow rule read throughput, we installed 3000 randomly generated flow rules into C ONOS/ODL controllers and transmitted a flow rule read request through the northbound REST API of a target controller. Upon receiving the read request, the target controller replies with the requested flow rule. Similar to the flow rule installation experiment, the flow rule read throughput values of ONOS and ODL are normalized by the flow rule read throughput values when C is 1.[3]

As shown in Fig. 4b, the throughput of ONOS is rarely affected by C for to the same reason as in the flow rule installation throughput experiment. On the other hand, the two cases described in the flow rule installation experiment are considered in ODL. For case 1, the flow rule read requests do not require any agreement among the replicas. Therefore, a flow read request can be locally processed by the leader replica, and the throughput remains constant even with the increase of C. Also, in case 2, it can be seen that the throughput is rarely affected by C as the flow rule read requests do not require any agreement among the replicas. Meanwhile, case 2 shows drastically reduced read throughput compared to case 1, which can be explained by the same reason as in the flow rule installation experiment in ODL.

In the controller failover experiment, we run three ONOS/ODL controllers and select one ONOS/ODL controller as a master controller of D devices. For the failover scenario, the master controller is intentionally turned down, and the elapsed time from the last heartbeat message of the failure detector running in the master controller to the time when a role reply message from the last orphan device is received by a new master controller is measured.

Figure 5a shows the effect of the number of devices, D, on the failover time when $\Phi = 8$. For both ONOS and ODL, as D increases, the number of orphan devices upon a failure increases and the number of role request messages to be sent increases. As a result, the failover time increases with the increase of D as shown in Fig. 5a.

Figure 5b demonstrates the effect of Φ on the failover time when $D = 5$. It can be seen that the failover time increases as Φ increases both in ONOS and ODL. This can be explained as follows. As Φ increases, the cluster in ODL and ONOS becomes more conservative in determining a controller failure. Therefore, the elapsed time between the failure event and the failure detection event is incremental to Φ, and the failover time for ONOS and ODL increases accordingly.

Conclusion

In this article, we discuss high availability and scalability issues in SDN, and analyze ONOS and ODL approaches. Experimental results demonstrate that:

1. The flow rule installation throughput of ODL is significantly affected by the cluster size.
2. There is a trade-off between ONOS and ODL in terms of the flow rule consistency and the flow rule installation throughput.
3. The controller failover time is dependent on the number of devices and the failure detection threshold.

As open challenge:

1. There is a trade-off between inconsistency of network states and performance of network services.
2. The controller assignment problem in large-scale WAN environments must be addressed, where latencies between controllers and switches are significant.
3. Stability analysis in large-scale SDNs with a few tens or hundreds of controllers in a cluster should be further investigated, and ONOS and ODL will evolve to address these challenges.

References

[1] D. Kreutz *et al.*, "Software-Defined Networking: A Comprehensive Survey," *Proc. IEEE*, vol. 103, no. 1, Jan. 2015, pp. 14–76.

[3] The read throughput values of ONOS/ODL when C is 1 are 13,946.7 and 20,813.8 flows/s, respectively.

[2] Security Architecture for Enterprise Network (SANE) project. http://yuba.stanford.edu/sane/.
[3] M. Casado *et al.*, "Ethane: Taking Control of the Enterprise," *ACM SIGCOMM Comp. Commun. Review*, vol. 37, no. 4, Oct. 2007, pp. 1–12.
[4] N. McKeown *et al.*, "Openflow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Comp. Commun. Review*, vol. 38, no. 2, Apr. 2008, pp. 69–74.
[5] M. T. Özsu and P. Valduriez, *Principles of Distributed Database Systems*, Prentice-Hall, 2007.
[6] A. Krishnamurthy, S. Chandrabose, and A. Gember-Jacobson, "Pratyaastha: An Efficient Elastic Distributed SDN Control Plane," *Proc. ACM SIGCOMM Wksp. Hot Topics in Software Defined Networking 2014*, Chicago, IL, Aug. 2014.
[7] D. Ongaro and J. Ousterhout, "In Search of an Understandable Consensus Algorithm," *Proc. USENIX Annual Technical Conf. 2014*, Philadelphia, PA, June 2014.
[8] F. Botelho *et al.*, "On the Feasibility of a Consistent and Fault-Tolerant Data Store for SDNs," *Proc. Euro. Wksp. Software Defined Networks 2013*, Berlin, Germany, Oct. 2013.
[9] F. Botelho *et al.*, "On the Design of Practical Fault-Tolerant SDN Controllers," *Proc. Euro. Wksp. Software Defined Networks 2014*, Budapest, Hungary, Sept. 2014.
[10] M. Obadia *et al.*, "Failover Mechanisms for Distributed SDN Controllers," *Proc. IEEE Int'l. Wksp. Network of the Future 2014*, Paris, France, Dec. 2014.
[11] A. Dixit *et al.*, "Towards an Elastic Distributed SDN Controller," *Proc. ACM Wksp. Hot Topics in Software-Defined Networking 2013*, Hong Kong, Aug. 2013.
[12] N. Hayashibara *et al.*, "The φ accrual Failure Detector," *Proc. IEEE Int'l. Symp. Reliable Distributed Systems 2004*, Florianpolis, Brazil, Oct. 2004.
[13] M. Bjorklund, "YANG — A Data Modeling Language for the Network Configuration Protocol (NETCONF)," IETF RFC 6020, Oct. 2010.
[14] A. Demers *et al.*, "Epidemic Algorithms for Replicated Database Maintenance," *Proc. ACM Symp. Principles of Distributed Computing 1987*, Vancouver, BC, Aug. 1987.

Biographies

Dongeun Suh [M] (fever1989@korea.ac.kr) received his B.S. degrees from Korea University, Seoul, in 2012. He is currently a Ph.D. student in the School of Electrical Engineering, Korea University. From 2012 to 2016, he received a scholarship from Samsung Electronics. His research interests include SDN/NFV/DTN and multimedia streaming.

Seokwon Jang (imsoboy2@korea.ac.kr) received his B.S. degree from Korea University in 2015. He is currently an M.S. and Ph.D. integrated course student in the School of Electrical Engineering, Korea University. His research interests include SDN/NFV, future Internet, and programmable networking.

Sol Han (hs1087@korea.ac.kr) received his B.S. degree from Korea University in 2015. He is currently an M.S. and Ph.D. integrated course student in the School of Electrical Engineering, Korea University. His research interests include SDN/NFV, future Internet, and programmable networking.

Sangheon Pack [SM] (shpack@korea.ac.kr) received his B.S. and Ph.D. degrees from Seoul National University, Korea, in 2000 and 2005, respectively, both in computer engineering. In 2007, he joined the faculty of Korea University, where he is currently a professor in the School of Electrical Engineering. He was the recipient of the Korean Institute of Communications and Information Sciences (KICS) Haedong Young Scholar Award 2013 and the IEEE ComSoc APB Outstanding Young Researcher Award in 2009. His research interests include future Internet, softwarized networking (SDN/NFV), mobility management, and mobile cloud networking/edge computing.

Myung-Sup Kim (tmskim@korea.ac.kr) received his B.S., M.S., and Ph.D. degrees in computer science and engineering from POSTECH, Korea, in 1998, 2000, and 2004, respectively. He joined Korea University in 2006, where he is currently a professor in the Department of Computer and Information Science. His research interests include Internet traffic monitoring and analysis, SDN/NFV, and Internet security.

Taehong Kim (taehongkim@cbnu.ac.kr) received his Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST) in 2012. He has been an assistant professor with the School of Information and Communication Engineering, Chungbuk National University, Korea, since March 2016. He worked as a research staff member with Samsung Electronics and ETRI from May 2012 to February 2016. His research interests include wireless sensor networks, the Internet of Things, and SDN/NFV.

Chang-Gyu Lim (human@etri.re.kr) is a senior engineer of SDN Research Section, ETRI, Korea. He received his Master's degree at KAIST in 2002. His key research interests are: future Internet, software defined networking, and transport networks.

Enabling Highly Dynamic Mobile Scenarios with Software Defined Networking

Alberto Huertas Celdrán, Manuel Gil Pérez, Félix J. García Clemente, and Gregorio Martínez Pérez

The authors present a mobility-aware and policy-based on-demand control network solution oriented to the SDN paradigm. This is in charge of managing, at runtime, the service and/or system state with high-level policies, which consider the mobility of users and services, the network statistics, and the infrastructure location.

This work has been supported by a Séneca Foundation grant within the Human Resources Researching Training Program 2014, the European Commission Horizon 2020 Programme under grant agreement number H2020-ICT-2014-2/671672 — SELFNET (*Framework for Self-Organized Network Management in Virtualized and Software Defined Networks*), the Spanish MICINN (project DHARMA, *Dynamic Heterogeneous Threats Risk Management and Assessment*, with code TIN2014-59023-C2-1-R), and the European Commission (FEDER/ERDF).

Digital Object Identifier: 10.1109/MCOM.2017.1600117CM

Abstract

Mobile devices have promoted users' mobility; therefore, there is a necessity to provide services that accomplish users' requirements at any place and time. With this, location becomes a key aspect of providing the dynamism required by solutions like the provisioning of reasonable mobile services by service provider networks. In that sense, the SDN paradigm arose to evolve from current static networks, which are manually configured by administrators, toward dynamic networks able to manage on their own at runtime and on demand. Solutions managing the SDN resources by using policies have been proposed, but they do not consider one of the main aspects to network dynamism: mobility. This article presents a mobility-aware and policy-based on-demand control network solution oriented to the SDN paradigm. This is in charge of managing, at runtime, the service and/or system state with high-level policies, which consider the mobility of users and services, the network statistics, and the infrastructure location. In this context, we define different use cases with the concerns of end users when they are in very crowded places, and the solutions provided by our solution through policies: balancing the network traffic between the infrastructure located close to the overloaded one; creating or dismantling geolocated virtual network infrastructure when the existing one is not enough, or is misused to meet end-user demands; and restricting specific network traffic in critical scenarios, like in sports events where crowds consume services with large bandwidth.

Introduction

The recent technology advancements in mobile devices and networks have encouraged users' mobility; thus, location is one of the most important aspects for knowing where devices, resources, and people are. Location information can provide useful evidence with which to develop new proposals and solutions. For example, the European Commission is making great efforts, funding the Horizon 2020 Programme to define new use cases where mobility and dynamism are key aspects. Under the Fifth Generation Public Private Partnership (5G-PPP) initiative, the EU project METIS-II [1] is proposing several use cases that highlight the provisioning of reasonable mobile broadband by service provider networks, with high levels of service experience in crowded areas (e.g., stadiums and shopping malls) and even with end users on the move (e.g., in cars and trains). Other initiatives are being conducted in parallel in other countries or continents, such as 4G Americas [2], where leading telecommunications service providers and manufacturers are fostering the advancement of the LTE mobile broadband technology and its evolution beyond to 5G, or the IMT-2020 (5G) Promotion Group [3], including main operators, vendors, and research institutes in China.

Managing the dynamism displayed by the previous proposals requires a deep change from current networks, where service provider administrators usually configure the network depending on triggered events, toward self-organizing networks (SONs) [4], which are able to monitor, manage, and configure on their own at runtime and, depending on different factors, among which location of users receiving a service is critical. This diversity requires that service provider networks collect and analyze large quantities of data, thereby increasing the network management complexity.

In order to ease network management, the software defined networking (SDN) paradigm arose [5]. SDN is a paradigm where a central software program, called the *controller*, is the brain of the network to manage its behavior, thereby making network devices become simple packet forwarding elements. This paradigm focuses on the separation of the *control plane* (where the controller is) from the *data plane* (where the forwarding devices are); the definition of a logically centralized controller; the use of open interfaces between the control and data planes; and the programmability of the network by applications. These features provide several benefits, such as ease of changing the network configuration through software rather than typing commands in network devices. Nowadays, we can find several solutions focused on deciding how the SDN resources have to be managed at runtime.

For example, NetGraph [6] provides a scalable graph library and interfaces with the controller to support network management functions, such as runtime monitoring and diagnostics. Another example is Procera [7], an event-driven network control framework that uses high-level policies to manage and configure the network state. This solution enables dynamic policies, which are translated into a set of forwarding rules to manage the network state by the controller. Following the policy-oriented approach, we find OpenSec [8]. Opensec is an OpenFlow-based framework

The authors are with the University of Murcia.

0163-6804/17/$25.00 © 2017 IEEE

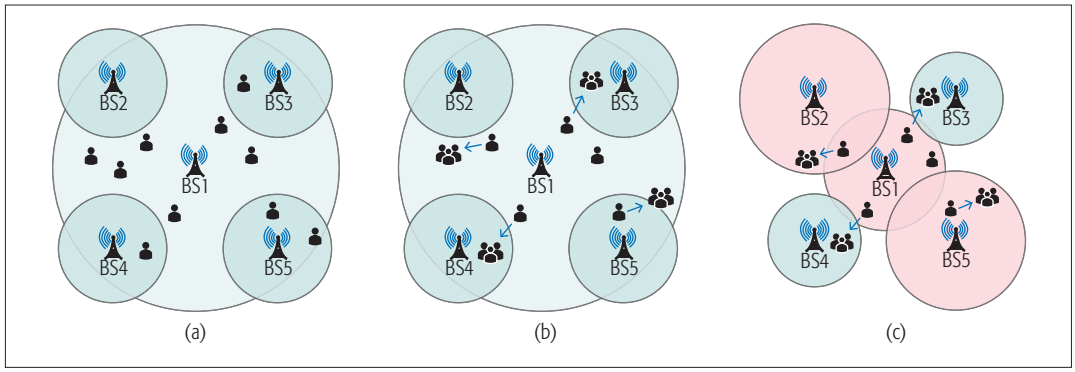


Figure 1. Network with enough resources providing low-quality services in a crowded scenario: a) initial scenario; b) considered scenario; c) managed scenario.

that allows network operators to describe security policies using human-readable language to implement them across the network.

Until this point, we have seen that there are solutions allowing the SDN controller to manage the network resources at runtime, using policies defined by service provider network administrators beforehand. However, these solutions do not consider one of the main aspects that provide network dynamism: mobility. We think that it is a must to consider users' mobility and the location of the network resources so as to manage and configure the SDN state in a more accurate way. In that sense, this article presents a mobility-aware and policy-based on-demand control network solution oriented to the SDN paradigm. Specifically, our solution is in charge of managing the SDN resources at runtime, using high-level policies that consider the mobility of users and services, the network statistics, and the infrastructure location. These policies are oriented to guarantee end users' experience in very crowded places (e.g., stadiums, shopping malls, and unexpected traffic jams). To this end, the policies decide when the SDN should balance the network traffic between the infrastructure located close to the congested one; when the SDN should create or dismantle physical or virtual infrastructure in case the congested one is not enough to meet the end-user demand; and when the SDN should restrict or limit specific services or network traffic in critical situations produced by large crowds using services in specific areas.

Use Cases in a Dynamic Mobile Scenario

This section shows a dynamic mobile scenario composed of four different use cases, with which to illustrate the service provisioning concerns that end users can find when they are in a very crowded place (e.g., open air festivals, traffic jams, stadiums, and public events with lots of people). The first use case shows a concern when the network provides low-quality services, even when having enough resources to meet end users' requirements. The second use case considers that the network does not have enough resources and provides low-quality services, whereas in the third use case the network does not have enough resources and is not able to provide services. In the fourth use case, the network misuses its resources to provide services. We explain in detail how our solution manages these concerns to ensure a good experience for end users.

A use case showing the first concern is shown in Fig. 1a, where a central base station (BS1) and four secondaries (BS2, BS3, BS4, and BS5) are located along a specific area. When large crowds are formed, and end users move across the networking area, BS1 is overloaded. Figure 1b shows this situation. BS1 is congested because it is providing services to a lot of users, and BS2 and BS5 just to a few. To solve it, our solution allows load balancing at runtime between the BSs located close to the congested one (BS1). In that sense, Fig. 1c shows how the zoom cell size load balancing technique [9] decreases the BS1 cell size and increases BS2 and BS5 cell sizes to ensure end users' experience. It is worth noting that when the crowd moves inside or outside the area, our system dynamically balances the load traffic, increasing or decreasing the size of the BSs' cells. An example of this situation could be an open festival with a central BS covering the whole festival, and four BSs close to the concert stages. Once the concerts start, the crowd moves to the concert stages and overload the central BS (e.g., sharing photos and videos through social networks).

Regarding the second concern, produced when the network does not have enough resources and provides low-quality services, Fig. 2a shows a use case where BS1 and four generic hardware (HW) elements with 3G/4G antennas are located along a specific area. In this context, Fig. 2b shows the moment when a mobile crowd is formed and the BS1 cannot meet the end users' requirements. To manage this situation, our proposal allows creating virtual BSs (BS2, BS3, BS4, and BS5) at runtime by using at will the generic hardware elements. Figure 2c depicts the situation managed by our solution. The created virtual BSs provide services once the network traffic is balanced. It is worth noting that once the crowd is gone, our proposal dismantles the virtual BSs, and the generic HW will be available to the service provider network. This situation is shown in Fig. 4, which is explained in detail at the end of this section. An example of this second use case could be a motorway with a BS and four generic HW elements located along its area. Due to weather conditions, a traffic jam is formed, and the BS cannot meet the requirements of the crowd, even knowing the atmospheric forecast. To solve it, our solution decides to create four virtual BSs from the existing generic HW and balances the traffic between them.

The use cases described earlier may become critical situations when the network does not have more available resources to meet the crowd's

Until this point, we have seen that there are solutions allowing the SDN controller to manage the network resources at run-time, using policies defined by service provider network administrators beforehand. Yet, these solutions do not consider one of the main aspects that provide network dynamism, i.e. mobility.

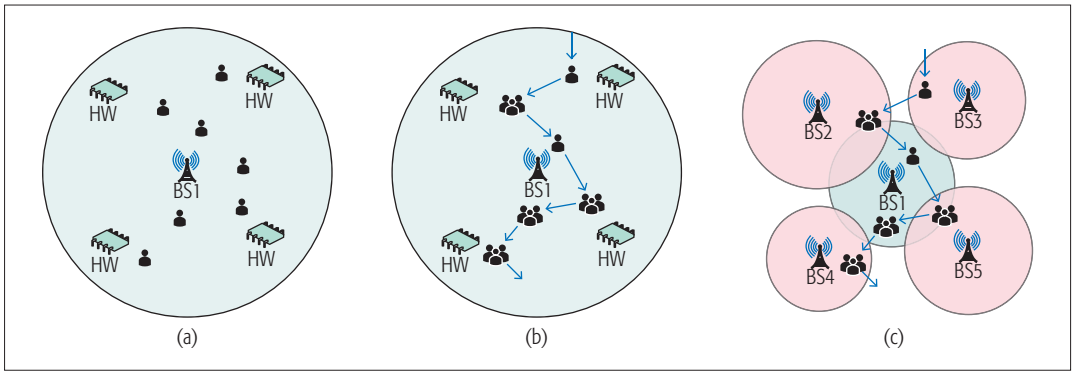


Figure 2. Network without enough resources providing low-quality services in a crowded scenario: a) initial scenario; b) considered scenario; c) managed scenario.

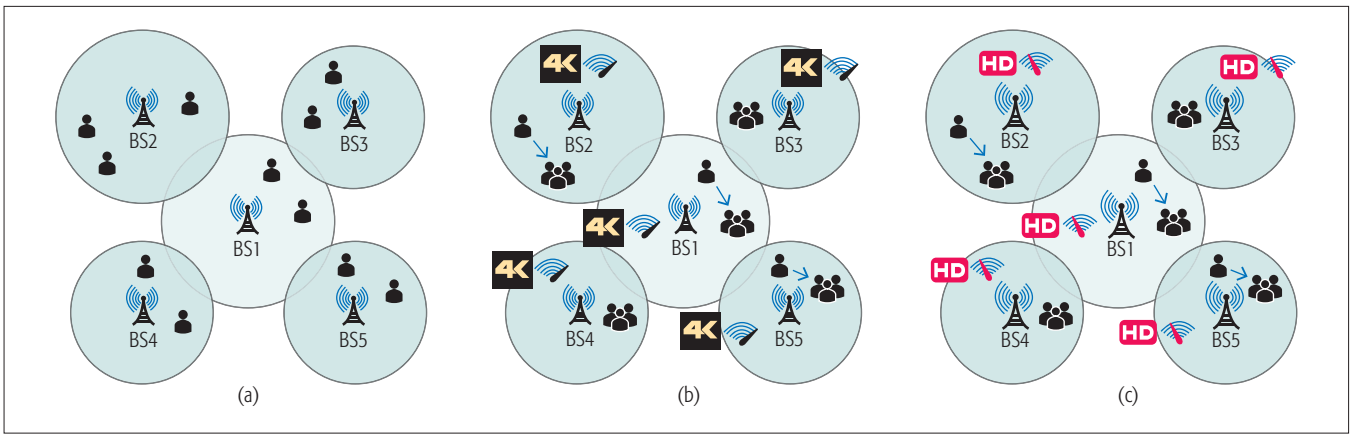


Figure 3. Network without more resources unable to provide services in a crowded scenario: a) initial scenario; b) considered scenario; c) managed scenario.

needs. In this sense, Fig. 3a shows a new use case where the whole available network infrastructure (all the BSs) is already deployed in a certain area to ensure the end users' experience. Figure 3b depicts how this situation could become critical, causing the network to not be able to provide services when more users come and consume services that require a large bandwidth such as 4K ultra high definition (UHD) video. To solve it, Fig. 3c shows the scenario where our solution decides that all the BSs reduce the quality of video service from 4K UHD to high definition (HD), and limits the bit rate to decrease the network congestion. As in the previous use cases, the reverse process (restrictions are removed) is performed when crowd conditions disappear. An example of this use case could be the Super Bowl, where the whole network infrastructure is deployed and balanced in the stadium. At the celebration, the crowd massively makes use of the network to send 4K-UHD videos, thus causing the BSs to be unable to accomplish the demand.

Up until now, we have seen several concerns generated when large crowds are formed. However, it is important to consider the reverse process, when the crowds are gone and the resources are not used efficiently, wasting energy resources. In that sense, the fourth concern arises when the network uses unnecessary resources to provide services. Figure 4a shows a use case where a physical BS (BS1) and four virtual BSs (from BS2 to BS5) provide services in a specific crowded area. Figure 4b shows the moment when the crowd starts leaving the area, and all BSs continue providing services to a few users. In order to prevent the misuse of resources, our proposal allows dismantling the virtual BSs at runtime. Figure 4c depicts this situation. The virtual BSs are dismantled, and BS1 provides services after increasing its cell size through load balancing.

Following with the traffic jam example, the jam begins clearing up when the weather conditions improve, and the virtual network infrastructure previously created is not necessary. In that case, our solution decides to dismantle the four virtual BSs and balance their traffic to BS1 by increasing its cell size to cover the whole motorway area.

SDN Management Policies

Policy-based management allows the simplification and automation of the network administration processes [10]. By using policies, the SDN paradigm can control the network state at runtime and on demand in order to guarantee the end users' experience. Among the different sets of policies, we emphasize here the use of mobility-aware management-oriented policies, defined by the service provider network administrator to decide the actions made by the SDN according to the network infrastructure statistics and location, and the mobility of users and services. In our solution, the schema of the rules shaping the policies are composed of the elements shown in Table 1, being

$$Type \wedge Resource \wedge Metric \wedge Location \wedge Date \rightarrow Result$$

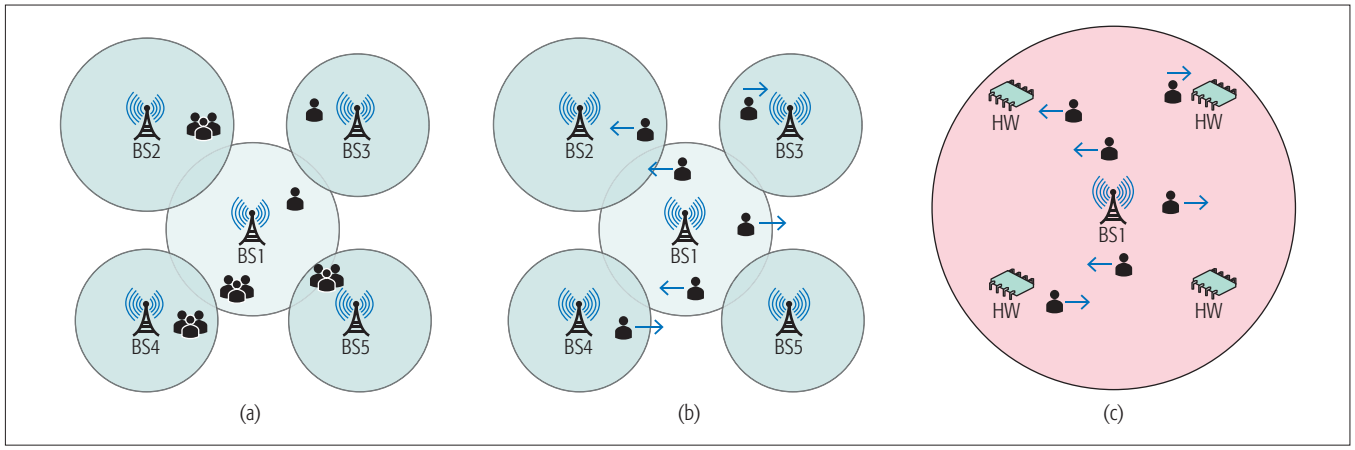


Figure 4. Network misusing resources to provide services: a) initial scenario; b) considered scenario; c) managed scenario.

Policies to Guarantee End Users' Experience

We introduce below the three kinds of policies required to manage the concerns depicted in the previous use cases, although other sorts of policies could be defined at will because the proposed solution herein presented is extensible.

Load Balancing Policies: These policies are in charge of deciding when, where, and why load balancing of the traffic between the network resources is needed, this being a key aspect in the SDN paradigm for managing and forwarding, at runtime, the packets passing through the network, considering their location, the date, and the metrics previously defined. These parameters are optional in this kind of policies. It is important to note that we are not proposing a new load balancing solution; we are able to use any load balancing solution.

Infrastructure Policies: These policies allow the SDN paradigm to create or dismantle virtual network resources located at specific locations. As with the previous kind of policies, they can be applied in a proactive way knowing when the network needs more infrastructure. As before, the *Date*, *Metric*, and *Location* parameters are also optional.

Restriction Policies: They manage the network or SDN to guarantee the end users' experience. These policies allow the SDN paradigm to disable or limit the traffic of given network resources or services in case the traffic overload is critical.

Managing the Dynamic Mobile Scenario

It is shown below how our solution manages the concerns presented earlier and how we guarantee end users' experience in very crowded places when important changes in the population occur in a short period of time.

Regarding the first concern, when the network has enough resources but provides low-quality services, our solution defines a generic load balancing policy. The policy defined below indicates, for example, that when the ABf value of any base station is within *Yellow* range values (the range of this alarm is set by the service provider administrator depending on the state and characteristics of the scenario), the network should try to balance the traffic load between the BSs located in the same area as the congested ones.

```
Type(#LoadBalancing) ∧ BaseStation(?bs) ∧
Location(?bs,?area) ∧ locatedBaseStation(?area,?nearBs) ∧
hasABf(?bs,?abf) ∧ inRange(?abf,#Yellow) →
balance(?bs,?nearBs)
```

In this policy, *BaseStation* is a possible value of the *Resource* element (defined in our policy schema as shown in Table 1); *Location* and *locatedBaseStation* are modeled by the *Location* element; *hasABf* is a specific *Metric*; and *balance* is a possible value of the *Result* element. Considering our open air festival scenario, Fig. 1c shows in red the changes made by this policy in the festival area.

To manage the second concern, when the network does not have enough resources and provides low-quality services, our solution defines an *Infrastructure* policy. As an example, the policy defined below creates new virtual BSs from generic hardware located close to the congested one when ANPPF of any BS is within *Orange* range values (this alarm is also defined by the service provider administrator, whose range of values is higher than *Yellow* range).

```
Type(#Infrastructure) ∧ BaseStation(?bs) ∧ Location(?bs,?area)
∧ locatedResources(?area,?resource) ∧ hasANPPF(?bs,?anppf)
∧ inRange(?anppf,#Orange) → create(?resource,#BaseStation)
```

In this policy, *BaseStation* is a value of the *Resource* element; *Location* and *locatedResources* are shaped by the *Location* element; *hasANPPF* is a kind of Metric; and *create* makes reference to a possible value of the *Result* element. Following the traffic jam scenario, Fig. 2c depicts in red the virtual BSs (BS2, BS3, BS4, and BS5) created from the existing generic hardware. Furthermore, a new load balancing policy is necessary once the virtual BSs are created, in order to balance the network traffic between them.

Regarding the third concern, when the network does not have more resources and it cannot provide services, our solution avoids this situation with two *Restriction* policies. The first one is in charge of disabling the 4K-UHD video traffic of the BSs located in the congested area. It is important to note that a disable action does not filter the video service, but disables a specific quality, and the service is provided with lower quality. Below we can find this policy.

```
Type(#Restriction) ∧ BaseStation(?bs) ∧ Location(?bs,?area)
∧ locatedBaseStation(?area,?nearBs) ∧
Service(?nearBs,?service) ∧ hasABf(?bs,?abf) ∧
inRange(?abf,#Red) → limit(?service,#BitRate)
```

The second *Restriction* policy limits the bit rate

Element	Values	Description
Type	Load balancing, infrastructure, restriction	Indicate the kind of policy
Resource	Base station, switch, service, intrusion detection system, etc.	Network element whose information is being managed
Metric	Average of bytes per flow (ABf), average number of packets per flow (ANPPF), average duration per flow (ADf), etc.	Define the term that encompasses the different parameters that can be used to evaluate the network state
Location	Geographic position, area, etc.	Position or region where the policy will be enforced
Date	Date, hour, timestamp, etc.	Moment or period of time at which the policy will be applied
Result	Balance, create, dismantle, disable, limit	Action performed over the network when the policy is applied

Table 1. Elements that compose the basis of our mobility-aware management policies.

of the services provided by the BSs located in the congested area.

$$\text{Type}(\#\text{Restriction}) \wedge \text{BaseStation}(?bs) \wedge \text{Location}(?bs,?area) \wedge \text{locatedBaseStation}(?area,?nearBs) \wedge \text{Service}(?nearBs,?service) \wedge \text{hasABf}(?bs,?abf) \wedge \text{inRange}(?abf,\#\text{Red}) \rightarrow \text{disable}(?service,\#\text{4K-UHDVideo})$$

In both policies, *BaseStation* and *Service* are values of the *Resource* element; *Location* and *locatedBaseStation* are modeled by the *Location element*; *hasABf* is a kind of *Metric;* and *disable* and *limit* are values of the *Result* element. Figure 3c depicts the Super Bowl event, where all BSs located at the stadium area decrease the video quality (from 4K-UHD to HD) and limit the bit rate.

Finally, the fourth concern arises when the crowd is gone and the network resources are misused. Our solution defines an *Infrastructure* policy that dismantles the misused virtual BSs located close to the underloaded one when the ANPPF value of any BS is less than *Yellow* range values.

$$\text{Type}(\#\text{Infrastructure}) \wedge \text{BaseStation}(?bs) \wedge \text{Location}(?bs,?area) \wedge \text{locatedBaseStation}(?area,?nearBs) \wedge \text{hasANPPF}(?bs,?anppf) \wedge \text{lessRange}(?anppf,\#\text{Yellow}) \wedge \text{hasANPPF}(?nearBs,?nearAnppf) \wedge \text{lessRange}(?nearAnppf,\#\text{Yellow}) \rightarrow \text{dismantle}(?nearBs,\#\text{BaseStation})$$

As before, *BaseStation* is a value of the Resource element; *Location* and *locatedBaseStation* are shaped by the *Location* element; *hasANPPF* is a kind of *Metric*; and *dismantle* corresponds to a value of the *Result element*. Following the traffic jam scenario, Fig. 4c shows the virtual BSs dismantled and converted again in generic HW. Furthermore, a new load balancing policy is necessary once the virtual BSs are dismantled in order to balance the network traffic to BS1.

Architecture

This section describes our mobility-aware architecture for managing networks oriented to the SDN paradigm at runtime and on demand. Figure 5 shows the proposed architecture, where the *SDN plane* contains the elements forming the layers of the SDN paradigm, and the SDN management plane depicts the components composing our solution.

SDN Plane

One of the main features of SDN is the decoupling of the control from the data plane. In that sense, our proposal has the data plane at the bottom layer, where physical and virtual network infrastructure (BSs, switches, routers, etc.) forwards and manipulates packets, not having any intelligent control. The networking logic control is allocated in the control plane, where the Controller component lies.

To exchange information between control and data planes, our solution makes use of OpenFlow [11]. This is one of the most common southbound SDN interfaces and allows our Controller to get statistical data about the network traffic, as well as the management of the network infrastructure through software. Nowadays, there are many OpenFlow-capable controllers, such as OpenDaylight, which is used by our solution.

Finally, the application layer is at the top of the SDN stack. This layer contains the applications that use the services provided by the Controller to perform tasks related to the network. Among the existing applications, we highlight three of them used in our solution. The network virtualization application is in charge of managing the virtual network resources by using a well-known open source software platform called OpenStack Networking (Neutron). Other solutions can be found in the literature such as FlowN [12], which presents an architecture for SDN virtualization. This allows tenants to specify their own address space, topology, and control logic. The second application is the load balancing application, which redistributes the network traffic between the network resources. On this topic, several solutions have been proposed, such as the one presented in [9], where load balancing is performed to increase or decrease the cell size according to the traffic load, user requirements, and network conditions. The last application is the service restriction application, which restricts the network traffic by considering different parameters, such as the bit rate, services, and ports.

SDN Management Plane

The main component of our solution is the policy engine. This component is in charge of making decisions over the SDN applications, considering network statistics, the infrastructure location information, and the network policies. Among the possible decisions, we highlight three of them. The first one consists on notifying the load balancing application about the need to redirect the traffic. The second one is focused on deciding if the network virtualization application has to create or dismantle virtual resources. The last decision is aimed at knowing if the service restriction application should limit or disable some kind of traffic.

To perform the previous decisions, the policy engine uses network policies, defined by the service provider network administrator, and geospatial network statistic information provided by the Collector. This component generates geospatial network statistics, by joining the information received from the Controller and the infrastructure location obtained from the location middleware. In order to deploy the Collector, we have

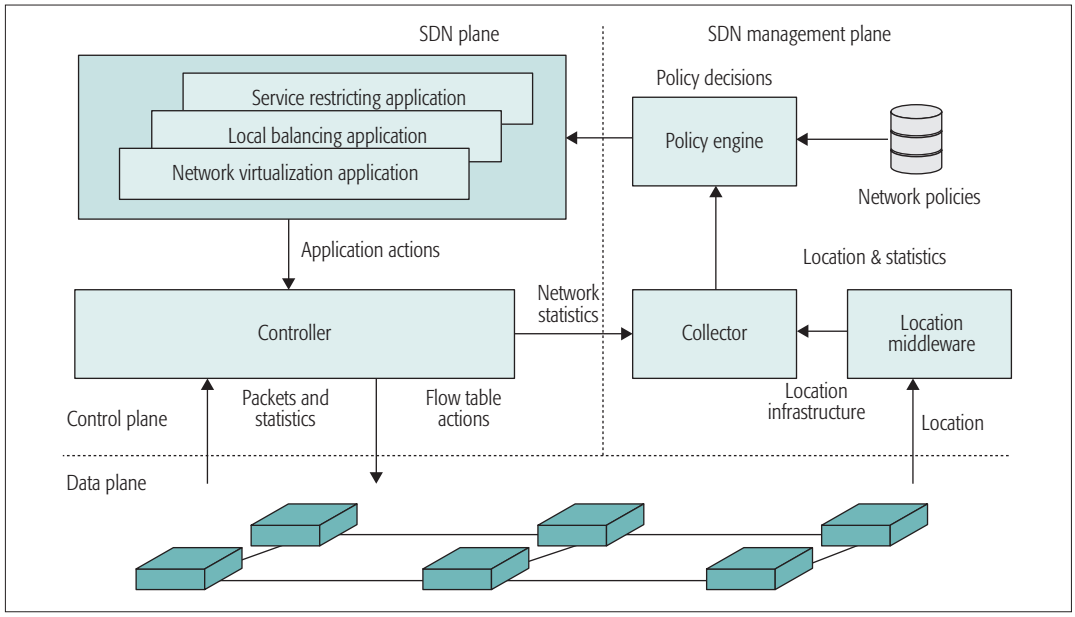


Figure 5. Architecture of the proposed mobility-aware and policy-based solution.

several options such as the extended version of IPFIX, which includes the location of the network infrastructure to generate network statistics.

Finally, the location middleware component obtains the locations of the network infrastructure. This is an independent middleware that provides independence to our solution with regard to the location system used, thus allowing the location middleware to choose the best location system or middleware depending on the environment.

Conclusion and Future Work

This article has presented a mobility-aware solution to manage, at runtime, networks oriented to the SDN paradigm, considering users' mobility as a key aspect for service provision. This proposal uses management policies to decide, on demand, the actions performed by the network, considering the mobility of users and services, the network statistics, and the infrastructure location. These policies ensure the end users' experience in crowded scenarios, balancing the network traffic between the infrastructure located close to the congested one, when the SDN has enough resources but provides low-quality services; creating virtual network infrastructure when the SDN does not have enough resources and provides low-quality services; and restricting specific network traffic when the SDN does not have more resources and is unable to provide services.

As next steps of research, we plan to validate our solution in a 5G advanced self-organizing network, as this has an important intelligence component oriented to the SDN paradigm. This scenario is proposed in the EU project for 5G called Selfnet, which is included in the 5G-PPP initiative and where the authors of this article are currently working.

References

[1] EC, EU Project METIS-II"; https://metis-ii.5g-ppp.eu.
[2] 4G Americas, "The Voice of 5G for the Americas"; http://www.4gamericas.org.
[3] "The Chinese IMT-2020 (5G) Promotion Group"; http://www.imt-2020.cn/en.
[4] H. Yang, X. Meng, and S. Lu, "Self-Organized Network-Layer Security in Mobile Ad Hoc Networks," *Proc. 1st ACM Wksp. Wireless Security*, Aug. 2002, pp. 11–20.
[5] R. Horvath, D. Nedbal, and M. Stieninger, "A Literature Review on Challenges and Effects of Software Defined Networking," *Procedia Computer Science*, vol. 64, 2015, pp. 552–61.
[6] R. Raghavendra, J. Lobo, and K.-W. Lee, "Dynamic Graph Query Primitives for SDN-Based Cloud Network Management," *Proc. 1st Wksp. Hot Topics in Software Defined Networks,* Aug. 2012, pp. 97–102.
[7] K. Hyojoon and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 114–19.
[8] A. Lara and B. Ramamurthy, "OpenSec: Policy-Based Security Using Software-Defined Networking," *IEEE Trans. Network and Service Management*, vol. 13, no. 1, Mar. 2016, pp. 30–42.
[9] N. Zhisheng *et al.*, "Cell Zooming for Cost-Efficient Green Cellular Networks," *IEEE Commun. Mag.*, vol. 48, no. 11, Nov. 2010, pp. 74–79.
[10] D. C. Verma, "Simplifying Network Administration Using Policy-Based Management," *IEEE Network*, vol. 16, no. 2, Mar. 2002, pp. 20–26.
[11] N. McKeown *et al.*, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Comp. Commun. Review*, vol. 38, no. 2, Mar. 2008, pp. 69–74.
[12] D. Drutskoy, E. Keller, and J. Rexford, "Scalable Network Virtualization in Software-Defined Networks," *IEEE Internet Computing*, vol. 17, no. 2, Mar. 2013, pp. 20–27.

Biographies

Alberto Huertas Celdrán (alberto.huertas@um.es) is a research associate in the Department of Information and Communication Engineering of the University of Murcia, Spain. His scientific interests include security, software-defined networking, semantic technology, and policy-based context-aware systems. He received M.Sc. and Ph.D. degrees in computer science from the University of Murcia.

Manuel Gil Pérez (mgilperez@um.es) is a research associate in the Department of Information and Communication Engineering of the University of Murcia. His scientific activity is mainly devoted to security infrastructures, trust management, and intrusion detection systems. He received M.Sc. and Ph.D. degrees in computer science from the University of Murcia.

Félix J. García Clemente (fgarcia@um.es) is an associate professor in the Department of Computer Engineering of the University of Murcia. His research interests include security and management of distributed communication networks. He received M.Sc. and Ph.D. degrees in computer science from the University of Murcia.

Gregorio Martínez Pérez (gregorio@um.es) is a full professor in the Department of Information and Communication Engineering of the University of Murcia. His research interests include security and management. He received M.Sc. and Ph.D. degrees in computer science from the University of Murcia.

Finally, the location middleware component obtains the locations of the network infrastructure. This is an independent middleware that provides independence to our solution with regard to the location system used, thus allowing the location middleware to choose the best location system or middleware depending on the environment.

DESIGN AND IMPLEMENTATION


Vijay K. Gurbani


Salvatore Loreto


Ravi Subramanyan

We access the world through our phones. At the risk of repeating ourselves: we access the world through our phones. Think about it — a device and a network that were once designed for a single application, voice, now transport mostly data, hence giving a new meaning to the phrase "data-driven network." This technical shift has consequences that have been subjects of research over the past two decades and will continue to be so in the future. The numbers behind the shift to data are simply immense, and as more money and subscribers continue to move toward accessing and exchanging data from their phones, the underlying network evolves to make data access and data exchange ever faster.

Pyramid Research forecasts that worldwide revenues for providing Internet access to residential customers in 2017 will be U.S. $188.7 billion and in 2021 the revenues are expected to top U.S. $222.7 billion [1]. In contrast, worldwide revenues for providing voice services to residential customers (including circuit-switched voice and packet-switched voice over a managed network) will decline from an estimate of U.S. $96.6 billion in 2017 to U.S. $86.1 billion in 2021. Clearly, this is a clarion call — if we needed one — to focus on moving data as expeditiously as we can over the service provider's network.

The two articles in this issue of the *IEEE Communications Magazine* Design and Implementation Series highlight the need and the mechanisms for ever faster data networks. In the first article, Cheng *et al.* note that telecommunications operators are eagerly identifying new sources of revenue from the network as revenues from voice calls decrease. They posit that exposing the network functionality as services that can be composed (or mashed up) into novel applications by users is a source of revenue. This is certainly not a new concept; we have been there before with TINA-C [2], SOAP [3], and Parlay-X [3]. However, the difference now is that the work of Cheng *et al.* allows users to compose services without having a computing background. Their work provides a drag-and-drop metaphor for composing novel services. To show the efficacy of their approach, they recruited 22 staff and student volunteers from a university and allowed them to get creative while composing services. Their insights from this experiment are noteworthy.

The second article, by Kim and Calin, moves from the user's point of view to the network's viewpoint. Specifically, they examine the performance of the split-TCP model prevalent in mobile networks today. Under such a model, the end-to-end TCP connection between a client and server (or a pair of peers) is broken down into two connections at the edge of the radio access network. A TCP connection is maintained between the mobile endpoint and a corresponding split-TCP host at the edge of the radio access network, and another TCP connection is maintained between the split-TCP host and the destination server (which is presumed to be in the core, or wired, network). Such an arrangement increases throughput and leads to faster error recovery for the entities on the radio access network. Kim *et al.* measure the performance of split-TCP over 3G and 4G LTE networks and discover impressive gains when split-TCP is used in 4G networks (3G networks exhibit modest gains). The lesson here is that as the access network speed increases, split-TCP shows more gains. This is important because advanced radio technologies like 5G promise to bring data rates of 1–10 Gb/s, leading to higher performance improvements on faster networks.

We welcome you to 2017 with our inaugural issue of the Design and Implementation Series. We continue to seek high-quality papers for the series that distill important lessons learned from ongoing research projects and forge new research ideas. We hope you will consider submitting papers to the series!

REFERENCES

[1] Pyramid Research, Global Fixed Forecast Pack, Nov. 4, 2016.
[2] TINA-C Service Architecture, v. 5.0, 1997; http://www.tinac.com, accessed Jan. 23, 2017.
[3] W3C SOAP Specifications, https://www.w3.org/TR/soap, accessed Jan. 23, 2017.
[3] ETSI OSA Parlay X, Parlay X 3.0 Specifications, https://docbox.etsi.org/zArchive/TISPAN/Open/OSA/ParlayX30.html, accessed Jan. 23, 2017.

BIOGRAPHIES

VIJAY K. GURBANI [M'98] (vijay.gurbani@nokia-bell-labs.com) is a Distinguished Member of Technical Staff at Bell Laboratories' End-to-End Mobile Network Research department in Nokia Networks. He holds a B.Sc. in computer science with a minor in mathematics and an M.Sc. in computer science, both from Bradley University; and a Ph.D. in computer science from Illinois Institute of Technology. His current work is focused on scalable analytic architectures and algorithms for autonomic 5G networks. His research has resulted in products that are used in national and international service provider networks. He has over 60 publications in peer-reviewed conferences and journals, five books, seven granted U.S. patents, and 19 Internet Engineering Task Force (IETF) RFCs.

SALVATORE LORETO [M'01, SM'09] (salvatore.loreto@ieee.org) works as strategic product manager within the Media business unit at Ericsson, Stockholm, Sweden. He has made contributions in Internet transport protocols (e.g., TCP, SCTP), signal protocols (e.g., SIP, XMPP), VoIP, IP-telephony convergence, conferencing over IP, 3GPP IP Multimedia Subsystem (IMS), HTTP, WebRTC, and web technologies. He is also a quite active contributor to the IETF, where he has co-authored several RFCs and has served as co-chair for several working groups. For the IEEE Communications Society, he serves as a Design and Implementation Series co-Editor and Associate Technical Editor for *IEEE Communications Magazine*. He received an M.S. degree in engineer computer science and a Ph.D. degree in computer networking from Napoli University in 1999 and 2006, respectively. In 2014 he graduated as an executive M.B.A. from SDA Bocconi in Italy.

RAVI SUBRAHMANYAN [SM '97] (ravi.subrahmanyan@ieee.org) received M.S. and Ph.D. degrees in electrical engineering from Duke University, a B.Tech. from IIT Bombay, and an M.B.A. from MIT. He has over 50 refereed journal articles and conference publications, and holds over 20 issued patents. He has worked on various aspects of telecommunications, including hardware design and system architectures for data and video transport. He is a synchronization expert, was an Editor for *IEEE Communications Magazine* Feature Topics on Synchronization in NG Networks and NG911, and was a presenter on the Comsoc Webinar on Next Gen Synchronization Networks. He has served on various IEEE GLOBECOM and ICC conference committees and on Comsoc's TAOS TC since 2008, and is a Technical Editor for *IEEE Communications Magazine*.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

HUMAN-DRIVEN EDGE COMPUTING AND COMMUNICATION

AIMS AND SCOPE

The vision of Edge Computing considers that tasks are not exclusively allocated on centralized Cloud platforms, but are distributed towards the edge of the network (as in the Internet-of-Things and Fog Computing paradigms), and transferred closer to the business operations via the Content Delivery Networks. The traditional gateway becomes a set-top-box machine, with additional computation and storage capabilities, where micro tasks can be offloaded first, instead of directly to the Cloud. Mobile Edge Computing can also be a more suitable approach to extract knowledge also from privacy sensitive data, which are not to be transferred to third party entities (global cloud operators) for processing. The proliferation of the networking connectivity and the progressive miniaturization of the computing devices have paved the way to the sensor networks and their success in the automation of the several monitoring & control applications. Such networks are built in an ad hoc manner and deployed in an unsupervised manner, without an a-priori design. The consequent availability of long-range communication means at certain nodes of those networks has enabled the possibility of the Internet connection of the sensor network, to make use of cloud-based services.

The new challenge addressed by this Feature Topic (FT) is how to put users in the loop so that they can retake control of their information. The massive proliferation of personal computing devices is opening new human-centered designs that blur the boundaries between man and machine.

In addition, Edge services are also used to exchange the data collected and processed within the context of the IoT towards external services and/or to visualize them through traditional browser by the users. Now, the frontier for the research on the data management is related to the so-called Edge Computation and Communication, consisting of an architecture of one or more collaborative multitude of computing nodes that are placed between the sensor networks and the cloud-based services. Such a mediating level is responsible for carrying out a substantial amount of data storage and processing to reduce the retrieval time and have more control over the data with respect to the Cloud-based services and to consume less resources and energy to reduce the workload. The interdependencies among those three different levels of storage and computing within an IoT solution are complex and determining at which data should be collocated and elaborated is demanding but not simple to handle. Such a complex situation is further exacerbated if we consider to achieve Quality-of-Service goals such as reliability, availability, security, mobility and energy efficiency, without compromising the correct behavior of the system and the service duration of the devices batteries. Moreover, the interconnection between the sensor networks and the upper level is not simple to be supported, in fact, falls within those situations where traditional Internet architectures fail to provide it effectively. This is because the sensor networks are deployed on hostile and challenging environments implying intermittent connectivity, a heterogeneous mix of nodes, frequent nodal churn, and widely varying network conditions.

The analysis of human activity and their interactions with physical and digital artefacts will also be extremely useful for closing the control loop of adaptive distributed systems. This may open a new research playground for distributed systems that adapt to user behaviours in different contexts, moving more and more to the network edge through devices such as the 5th Generation mobile networks or 5th Generation wireless systems. The second aspect of the frontier of the current research is therefore related to the application of challenging networking solutions to support the Fog Communication and Computation in the Internet of Things.

Topics of interest include, but are not limited to::

- Novel models and architectures of Edge-centric computing
- Fog-to-Cloud integration and protocols
- Communication protocols and issues
- Crowdsensing and crowdsourcing information
- Human-driven design and implementation of edge computing
- Novel socially-informed architectures
- Delay-tolerant networks, opportunistic communication and computing
- Reliability and availability, mobility and connectivity in edge-centric computing
- User-guided management of Fog systems and services
- Resource management and provision
- Data harvesting and analytics in challenged networking
- Information centric and content-centric networking
- Distributed storage services
- Heterogeneity of edge systems
- Energy-efficient communication and computation
- Security and privacy, attacks and resiliency
- Secure and sensitivity-aware applications
- Novel safe methods for including humans in the data-analysis loop
- QoS-aware communication protocols
- Daily use applications and programming models
- Test and simulation tools for evaluating challenged systems
- Modelling and simulations

SUBMISSIONS

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions). Figures and tables should be limited to a combined total of six. The number of references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed if well-justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please submit a PDF (preferred) or MS WORD-formatted manuscript via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to the Author Center. Follow the instructions there. Select "November 2017/Human-Driven Edge Computing and Communication" as the Feature Topic category for your submission..

IMPORTANT DATES

Manuscript Submissions Due: April 30, 2017

Decision Notification: July 1, 2017

Final Manuscripts Due: August 15, 2017

Publication: November 2017

GUEST EDITORS

Florin Pop
University Politehnica of Bucharest, Romania
e-mail: florin.pop@cs.pub.ro

Aniello Castiglione
University of Salerno, Italy
e-mail: castiglione@ieee.org

Jiannong Cao
The Hong Kong Polytechnic Univ., Hong Kong
e-mail: csjcao@comp.polyu.edu.hk

Giovanni Motta
Google Inc., USA
e-mail: giovannimotta@google.com

Yang Yanjiang
Huawei Singapore Research Centre
e-mail: yang.yanjiang@huawei.com

Wanlei Zhou
Deakin University, Australia
e-mail: wanlei.zhou@deakin.edu.au

LSMP: A Lightweight Service Mashup Platform for Ordinary Users

Bo Cheng, Zhongyi Zhai, Shuai Zhao, and Junliang Chen

The authors propose a novel LSMP, which includes the SDM, SRM, SCM, and SEM. The SDM defines a unified structure for atom-services to reduce the development complexity. The SRM defines a structurized data flow to reduce the coupling of data transfer between atom-services. The SCM provides a design method for the service process, with key data as the driving characteristic.

ABSTRACT

With the telecom market reaching saturation in many regions and revenues from voice calls decreasing, telecom operators are attempting to identify new sources of revenue. For this purpose, these operators can take advantage of their core functionalities, such as location and call control, by exposing them as services to be composed by developers with third-party offerings available over the web. Mashups enable technology to compose those services into new applications. However, existing mashup tools have some limitations in such convergent web-service-based integration, such as extra programming efforts for ordinary users. To address these limitations, we propose a novel LSMP, which includes the SDM, SRM, SCM, and SEM. The SDM defines a unified structure for atom-services to reduce the development complexity. The SRM defines a structurized data flow to reduce the coupling of data transfer between atom-services. The SCM provides a design method for the service process, with key data as the driving characteristic. The SEM provides a container for the service logic execution. With these models, ordinary users can create new personalized services freely via an intuitive and easy-to-use UI on a web browser.

INTRODUCTION

The telecom business model is evolving. With the market reaching saturation and revenues from voice calls decreasing rapidly, telecom operators are aggressively looking at newer sources of revenue. Networks, services, and content are converging increasingly rapidly [1–3]. With the advent of Web 2.0 and the efforts of service providers, the number of open application programming interfaces (APIs) from Internet, telecommunications, and third-party services is growing rapidly. Mashup is emerging as a driving technology for lightweight service creation in the Web 2.0 era, which can enable end users to combine services to create new applications [4, 5]. Telcos have changed from isolated, monolithic legacy stovepipes to much more modular, Internet-style frameworks, opening up telecommunication features, such as voice calls, presence, messaging, contact management, and Session Initiation Protocol to the public. There is an ongoing effort to mix telecommunication features and web capabilities to obtain a more open and user-centric environment. For example, some approaches have added voice capabilities to their Web mashup, and several of Apple's iPhone APIs follow a form of mixed web and telco services creation. Users can combine various types of web services from the growing numbers of telco, Internet, and third-party services to achieve their goals under dynamic and complicated situations. The availability of open APIs enables ordinary users to generate new converged services more easily without excessive consideration of the underlying infrastructure and mechanisms required. Through a highly intuitive service creation environment, users can create a new service by composing previously existing open APIs in a more personalized manner. Eventually, the service mashup platform will help to create the personalized application in next-generation services by allowing the end users.

This article focuses on the design and implementation of a lightweight services mashup platform for user-centric service creation, the execution of which provides the tools necessary to allow everyone to build his/her own communication services without having specific background knowledge in computing. The rest of the article is organized as follows. We provide the requirements and review the related work. The novel architecture is then presented. We describe the implementation of our prototype, including the performance measurement and analysis. Lessons learned are discussed, and the conclusion and future work are covered in the last section.

MOTIVATION CASE

A smart parking lots service is a complex process composed of several hardware devices able to detect the occupancy level of parking spaces and software components integrated to manage the allocation and charging of these parking spaces by redirecting cars accordingly. Usually, such services are created by combining various types of web services from the telecom, Internet, and third-party services to assist motorists in the localization of available parking spaces so that they can decide on which space to select according to their own needs. The smart parking lots scenario is shown in Fig. 1.

Tony drives to his meeting in an unfamiliar place. When arriving at the scheduled place, he quickly locates his current position, searches nearby parking lots via a location-based service, and reserves a suitable parking spot via a mobile app. Meanwhile, the mobile geographic information system (GIS) provides a planning path to direct

Digital Object Identifier:
10.1109/MCOM.2017.1600309

The authors are with Beijing University of Posts and Telecommunications.

0163-6804/17/$25.00 © 2017 IEEE

him to the selected parking lot. As he enters the parking lot, the monitor at the entrance to the parking lot identifies the license plate automatically, and then launches the intelligent barrier and begins to count the service time. Tony then attends his meeting after parking his car. During the meeting time, his smartphone indicates the waiting time and provides some promotional information regarding parking. Tony uses his mobile phone to trigger a washing service in the parking lot to clean and polish his car while he is in the meeting. As he is leaving, he starts the charging service, and then the parking lot management system generates an electronic bill shown in his smartphone, which can quickly be paid using a mobile e-wallet. When the car arrives at the egress, the intelligent barrier lifts automatically to facilitate leaving.

Requirements and Related Works

Requirements

Enabling ordinary users to develop their own applications or compose application programs by combining different pieces available online in the form of public web services, APIs or data in various forms require simplifying current ordinary user development practices. Enabling the ordinary user to develop his/her own mashups requires an intuitive and easy-to-use user interface (UI) for building mashup-based compositions. The functionalities must also be provided through the building blocks in the mashup design environment, which requires an intuitive modeling construct capability. In other words, higher abstraction levels will result in more intuitive service creation tools. During runtime, immediate execution of the resulting mashup and the exchange and flow of data between components are required, and ordinary users need not be aware of what is happening behind the scenes.

Related Works and Discussions

Programming-based service mashup is based on SOAP-based web services or RESTful web services, which are satisfied to programmers' needs [6]. Thus, they do not comply with the above requirements for ordinary users. Their main limitation is that they are based on programming APIs, and hence are difficult and complex for ordinary users.

Some attempts to facilitate the service creation process based on SOAP-based web services have nonetheless been made in the business area. A typical approach is to rely on the service composition language, such as the Business Process Execution Language (BPEL) [7], which defines the control flow of structured and atomic activities and the partners and partner links for various web services to be included in the process. In addition, BPEL also supports asynchronous message processing, exception handling, transaction processing, and business life cycle management, and the BPEL process itself can be published as a web service. Although the BPLE languages facilitate the creation process, they are not appropriate for use by ordinary users. First, the business logic operators and data flow are too complex. Second, it is difficult for ordinary users to understand different computing concepts, such as flowcharts, inputs, and outputs, to create new services.

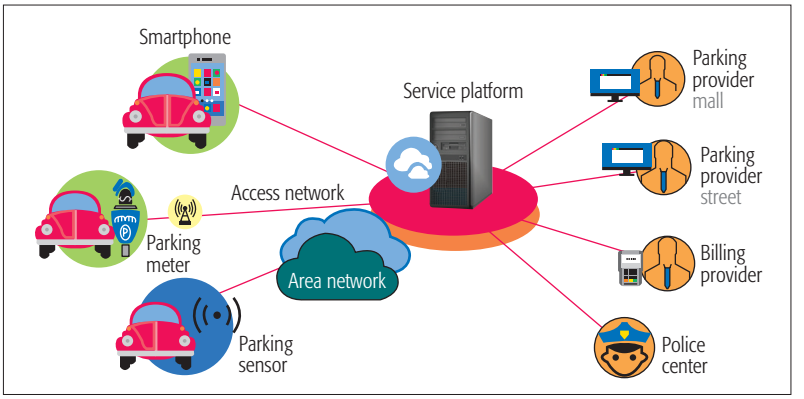


Figure 1. Smart parking lots service scenario.

Currently, ordinary users can develop their own applications by following mashup- or widget-based models [8, 9]. In the former case, ordinary users combine existing services and web feeds from multiple sources into a single web-based application using a specialized mashup editor. The widget-based model does not support any interaction between services offered by different service providers. Two important limitations should be highlighted regarding these tools. First, the flowchart basis is not intuitive enough for ordinary users. Concepts such as mapping the outputs of some services and inputs of others, loops, conditions, and regular expressions are not understood by most ordinary users. Second, although the mashup creation process is supported with an advanced graphical user interface (GUI), the service type is usually limited to a set of map and received signal strength (RSS) list patterns.

The Lightweight Telco Service Mashup Platform for ordinary-user-centric service creation and execution environments attempts to allow everyone to build his/her own communication services without having specific background knowledge in computing, which could facilitate the creation of successful applications accommodating end users' diverse needs, interests, and activities . The novelty of our work is introduced as follows:

First, we propose a novel service creation environment (SCE) for end users that is based on service-oriented architecture (SOA) principles, which can wrap complex software features within well-defined and reusable interfaces made available to third parties that can provide representations of services as visual Iframe objects which abstract from technical details, for example, their programmatic interface or communication protocol.

Second, ordinary users can construct composite services automatically via dragging and dropping the relevant service icons, optionally defining the flow-based programing (FBP)-style service composition with explicit data flow, which significantly enhances the intuitiveness of mashup creation for ordinary users without the need to understand programming concepts.

Third, to further enhance ordinary users' perceptions of the effects that individual services have on the final mashup applications in the service execution environment (SEE), which enables ordinary users to understand the current state and

Our platform has provided some type conversion modes that can facilitate coordination of the interaction of heterogeneous services smoothly. In addition to the syntactic representation, the attribute semantics can be annotated simultaneously by following the naming convention.

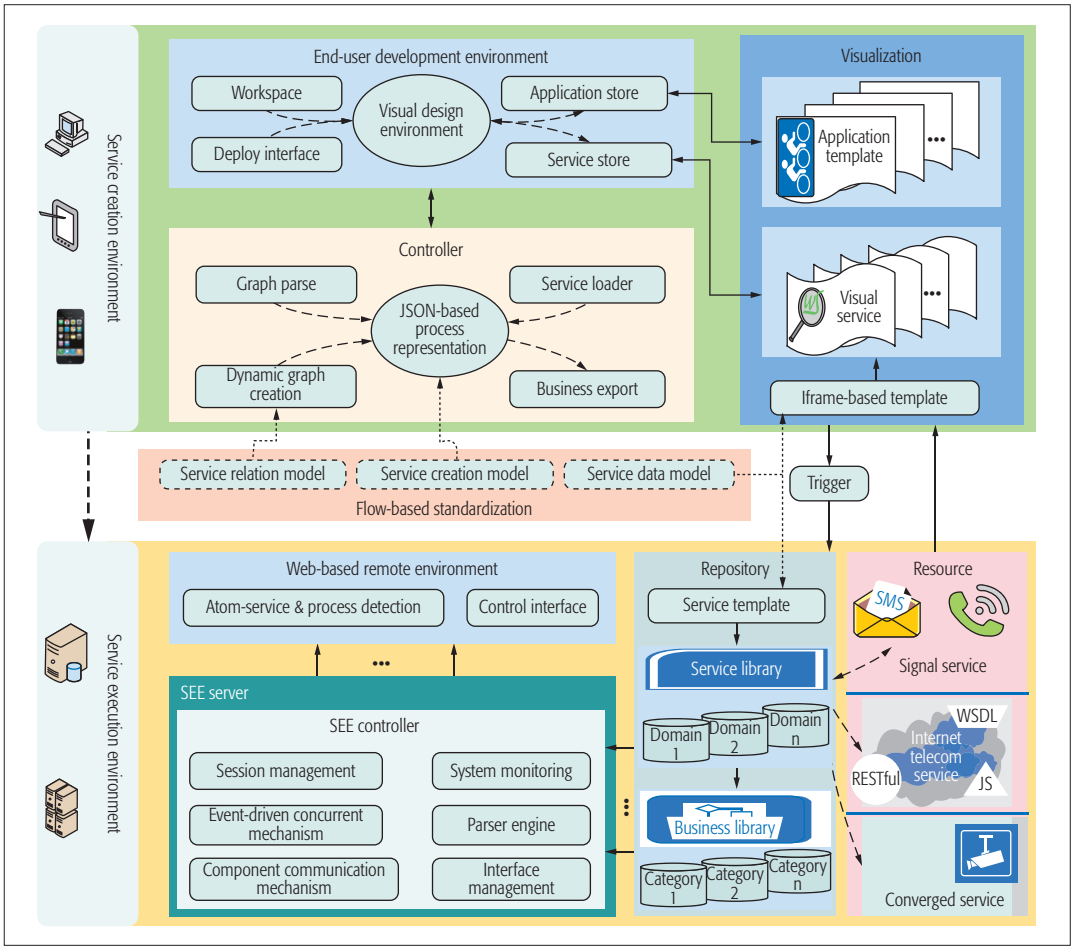


Figure 2. Lightweight Services Mashup Platform architecture, which contains the SCE and SEE.

feel of their composition, it is highly desirable to provide immediate visual feedback for any composition action and to support the immediate execution of the resulting mashup.

Finally, we implement a lightweight SCE environment based on currently popular browsers that is easy to use for ordinary users.

System Architecture

The LSMP includes the SCE and SEE. The SCE is associated with web browsers to develop new services, while the SEE takes charge of the execution of services and processes. Figure 2 shows the LSMP hierarchical architecture.

SCE

A browser-based development interface is implemented to satisfy multiple platforms, such as PCs and mobile phones. It mainly includes four parts: service store, application store, deploy interface, and workspace. Workspace is the area for drawing the service processes. The service store is used to display the basic feeds for assembling new services, in which each icon, corresponding to a virtual atom-service, can easily be dragged and dropped in the workspace. The application store can be used to show the service process templates that have been developed by ordinary users. The deploy interface can be used to deploy and push the developed service processes into the runtime environment.

SEE

The SEE controller is the basis for taking charge of the managing execution engine, including session management, system monitoring, a concurrent mechanism, a communication mechanism, and interface management. In the repository, the service library includes all the executable services, and the business library includes all the developed business processes. To increase flexibility, the atom-services in the library should be constructed from a multi-source resource, including native services for a database system, standardized services for third-party resources (WSDL, RESTful, and JS interface), and legacy systems.

Specifically, we present a series of standardized models in LSMP, including the service data model (SDM), service relation model (SRM), service creation model (SCM), and service execution model (SEM), to facilitate the service mashup in the flow-based approach. The SDM defines a standardized data-oriented service structure to facilitate the development and adaption of heterogeneous services. The SRM defines a basic data relation between two atom-services. The combination of SRMs is an explicit data flow graph that allows ordinary users to develop services intuitively. The SCM defines a data flow graph based on "data" and "flow," in which "data" can be extracted from requirements, and the "flow" should satisfy the constraints of the SRM.

SDM

Lightweight service mashup is a data-centered development process, in which atom-services, as basic units of data, concentrate on the service description, manipulation, and visualization of data. A SDM is defined by a quintuple:

SDM= <SD, AS, T, V, E>,

where SD is a text description of the service function and usage.

AS describes the attribute set of a service, which are data elements used in service. In the SDM, the data type of each attribute can be either simple type (e.g., *integer*, *string*, *JSON*) or complex types (e.g., *image*, *audio*, *custom type*). Our platform has provided some type conversion modes that can facilitate coordinating the interaction of heterogeneous services smoothly. In addition to the syntactic representation, the attribute semantics can be annotated simultaneously by following the naming convention.

T is used to transform one service input data into the desired outputs and can be defined by

$$T : A_I \xrightarrow{S_P} A_O,$$

which is a transformation that maps input set A_I to output set A_O in A_S; S_P is an operation set. Each operation is a functionality that can be obtained by the open service interface, for example, RESTful and SOAP APIs.

V describes the visual elements of the SDM (i.e., the graphical component). The graphical appearance of services provides an easy-to-use form for designing the business logics of mashup services. The visualization of services involves two parts: the graphical service framework and the visual element of service functions. In our platform, a widget-based representation has been constructed to display the service framework uniformly. For the service function, the service provider needs to specifically provide some operations to display the generated data.

E describes the extension of the SDM, the core of which is the extension of transformation *T*.

The SDM provides an abstract structure for describing and coordinating heterogeneous service in a unified way. Such a component model facilitates specifying the interoperability and data exchange between services in the high-level specification. In addition to this, an operable component needs to be constructed to support the third-party service access and the visualization of services.

Here, we introduce an IFrame-based implementation of the SDM by using web front-end technology, such as HTML5 and JavaScript. An IFrame is an HTML document embedded inside another HTML document on a webpage. The Iframe is often used to insert content from another source (e.g., a web-based service) into a web page. Although an Iframe behaves like an inline image, it can be configured with its own controls independent of the surrounding page's contents. Therefore, Iframe is fully suitable for being a visualization template of the SDM. Not only can it be constructed and modified by service providers easily, but it can also be used to display the natural characteristics of an SDM instance graphically. We developed a visualization template to describe the SDM with IFrame, as shown in Fig. 3. Because the Iframe instance is quite verbose, we use a comprehensive template to show how it can be developed. Some comments have been added to this template to show how each piece of code corresponds to the element in the SDM. At the beginning of the Iframe template, a dependency library for Iframe should be invoked by the service provider. When a dependency library is selected, the SDM instance will make an initial visual appearance; here, the pre-categorized dependency library has been prepared in advance to reduce the burden on the service provider. The IFrame service template provides a unified representation, which is useful for RESTful and JS services that do not have standard descriptions. Based on such template, end users can conveniently access some simple third-party services in our service library.

```
<html>
<head>
    <script citing dependency library to imeplement Iframe in LSCE,
      // i.e., implementation of S_F > </script>
    <script> definition of Inputs, Outputs & Transformation for atom-services, :
      LSCE.setInfo({
          basic information of atom-service //i.e., implementation of S_D
      }).addInputs({   //i.e., implementation of A_I and T
        definition of Input name & type;
        definition of execution action corresponding to input name. It will be
          triggered as soon as data arrives.
          – invoking service operation to manipulate data.;
          – invoking output port to transmit data;
      }).addOutputs ({    // i.e., implementation of A_O
          definition of Output name & type;
      }) </script>
    <script> user-defined inline operation to manipulate data.
      // i.e., implementation of E </script>
    <script> scripts to render Iframe page // i.e., implementation of V_P </script>
</head>
<body> web page contents of atom-service </body>
</html>
```

Figure 3. Visualization template describing the SDM with Iframe.

SRM

The SRM, a structurized explicit data flow, can describe the business logic in a natural way. An SRM is defined by a tuple:

SRM = <*f*, M_E, M_S>,

where *f* describes a data flow that transmits data from a service output attribute to a service input attribute. M_E indicates that data flow *f* must be grammatical matching between the attributes, that is, the data type of output end in *f* can be accepted by the type of input end. In the platform, some dependency relationship sets have been constructed for the data types. If the attributes of the data flow have a dependency relationship, we argue that the data flow is grammatical matching. When composing the services via the SRM, the system will automatically compute whether the data flow is grammatical matching.

M_S represents that data flow *f* must be semantic matching between the attributes, that is, the data semantics of output end in *f* can be accepted by the input end. In the platform, a seman-

```
path.wire{
    handler: function ( e ) {
    isAttribute: false
    lineNumber: 4298
    listenerBody: "function ( e ) {
      //Discard the second event of a jQuery.event.trigger() and
      //when an event is called after a page has unloded
    return typeof jQuerey !==core_strundefined
    && (!e || jQuery.event.triggered !==e.type) ?
    jQuery.event.dispatch.apply (eventHandle.elem, argumetns ) : undefined; }"
    node: path
    sourceName: "file:///.libs/jquery.js"
    type: "click"
    useCapture: false
}
```

Figure 4. An operation definition for the SRM.

tic tree of service attribute concepts have been constructed; thus, the semantic relationship of two attributes can be obtained from this tree. Similarly, the system can automatically check whether the data flow is semantic matching.

The SRM is actually a constraining data flow by which services can communicate with each other more reliably. Data flow graphs composed by SRMs can be used instead of the business logic (as control flow) between services; their expressive ability is described in detail in the next section. Compared to a conventional data flow, such flows specified by the SRM are structured objects that intuitively represent entity relations in the outside world. Thus, one can quickly describe manual applications with these SRMs. Not only can an SRM support the capability of distributed, heterogeneous applications, but it also facilitates the development of applications in a natural manner. The SRM provides an implicit control pattern that is relatively usable by ordinary users, and its expressive power has been proven to be reliable through our previous work [10]. Moreover, our approach has also been demonstrated to be feasible by an analysis of the *control flow patterns* (an industry standard in the *Workflow Patterns Initiative*) and are supported to illustrate the expressive power of the business process.

Because the SRM is a highly visual notation, a graphical representation will make the service development even more usable. We introduce a scalable vector graphics (SVG) implementation to render the SRM into a pipeline. Utilizing these graphical pipelines, ordinary users can describe an application as a graph in a natural manner, and the connection between the graph and the system procedure can be visualized seamlessly. Moreover, a series of operations related to an SRM must be implemented to control the graph drawing. Figure 4 gives an example of SRM operation, "path.wire." When the "click" event is captured, the path.wire will be executed to automatically create a pipeline according to the mouse position.

SCM

The SCM is used to specify the service creation process from requirements to data flow graphs. SCM is defined by a quad:

$$SCM=<D_R, S_R, SD_G, SRM_s>,$$

where D_R describes the related data to develop a mashup service. S_R describes the alternative services that might be used by the mashup service. SD_G describes a service dependency graph that implies the interaction relationships of services in the mashup. SRM_s describes a data flow graph of the mashup service, which is the outcome of the SDM.

The SCM provides a data-driven development pattern that can help end users create mashup applications. It is actually a graph creation process that relies on service attributes (i.e., data) and service dependencies. The following are the specific steps in the generation of a mashup service.

Step 1: According to the requirement for a mashup service, the end user needs to analyze the related data entities for the mashup initially. If there are some new services selected for SR in step 2, the developer can also extract related middle data entities from their service attributes iteratively.

Step 2: The end user should search and add the services according to the new data entities in step 1. If an added service is selected by the middle data entity of another service, the dependency relationship between these two services needs to be established.

Step 3: The *service dependency* graph needs to be checked to see whether it forms a complete dependency pathway from the initial services to the final services, by means of end users' domain knowledge or real-world experience. If there are some missing service attributes in the dependency graph, corresponding data entities will be added in the DR, and the end user needs to repair the dependency graph from step 1 again.

Step 4: According to the dependency graph, the user should complete all the data connections among services and simulate the execution effect of this mashup service. If the data flow graph works well, its service script will be deployed into the execution environment.

Through the above steps, ordinary users can develop a new service in a natural way. In the whole process, the focus is on the data (i.e., service attributes), which are easily understood and acquired by ordinary users. Thus, ordinary users can also construct service dependencies through their domain experience. Clearly, the creation method of the SCM is more loosely coupled as a result.

To make the connection between the data flow graph (i.e. SRM combinations) and the system procedure seamlessly, a service logic script corresponding to the data flow graph is also required that can be directly executed by machines. Here, we choose JSON as the service logic script of SRM combinations, and thus, the expression of service logic can be rapidly parsed by a core kernel for JavaScript.

SEM

The SCM describes how different atomic service instances interact with each other to create a higher level service using event-triggered patterns. The service logic of the SCM is specified by the set of connections between events (i.e., service attributes) and components, together with the Initial and Final Events. The execution of the functionalities of an atomic service is the responsibility

of the base service implementation, whereas the execution of the service logic is the responsibility of the service logic engine. We introduce a service execution model as the core of the service logic engine. The SEM is defined as follows:

$$SEM = <SCM, Event_{atomic}, Event_{composited}>,$$

where SCM is described in the JSON format. $Event_{atomic}$ is the atomic event, which means a single event is received by an application at a specific time, and can be defined as $e_p = e_p$ (*id*, *a*, *t*, *l*), where *id* is the primitive event's unique identifier; $a = \{a_1, a_2, , a_m\}$, $m > 0$, is the primitive event's attribute set, *t* is the primitive event's occurrence time, and *l* is the primitive event's location of occurrence. $Event_{composited}$ is the composited event, which means an event set that includes several atomic events with specified combination relation, and received from multiple events deriving from the application or passed between service primitives during application execution. It can be defined as $e_c = e_c$ (*id*, *a*, *c*, t_s, t_e), $t_e >= t_s$, where $c = \{e_1, e_2, , e_n\}$, $n > 0$, is an event set that includes several primitive events or composite events whose elements collectively constitute the composite event; t_s and t_e are the starting and ending times of the composite event, respectively.

Different events can be combined to generate a composite event. A composite event inherits all characteristics of its source events, and its occurrence time is also determined by the source events' occurrence times. This combination of events is defined by the event combination relation as follows.

The service logic engine exposes one endpoint per service composition on which it receives all the event notifications. Once a notification has been correlated with the service instance to which it belongs, the service logic engine will process the event according to the appropriate event-triggered pattern and will invoke one or more components on the atomic service execution containers. The service engine is implemented using Node.js, which makes it possible to support high concurrency; the events will be dispatched to a related Event Handler according to the dispatching strategy. The Event Handler only processes the related event logic, unless the Event Handler must process some extra logic that has a linkage effect on others, such as message passing or even launching a process workflow with multiple nodes. The core idea of the service execution container is based on the FBP model, and the complex business process is split into separate processing units that consist of components with specific functions. Thus, the business process can be viewed as a network composed of components. Components are connected by edges that pass data between them. Data will be processed by each processing unit when passing through the network and finally provide a result. Components communicate with each other by ports. A connection is attached to a component by means of a port. The communicating procedure will be triggered according to the event-driven mechanism. The initial message trigger will receive this message. This message may come from the Event Handler in the Event Broker or passed by a protocol, such as HTTP or WebSocket, outside the FBP Execution container. For compatibility, the JSON data format is used to describe the network. The network is divided into two parts: the component set and the connection set. The component set contains information such as a component's name and path. The connection set defines the connectivity between components.

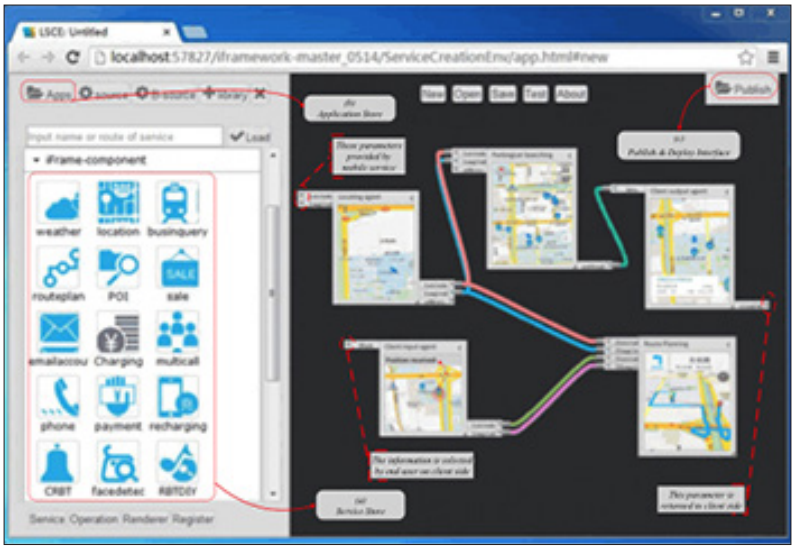


Figure 5. Service mashup environment and SPL demonstration.

Implementation and Illustration

This section shows how an ordinary user can create the useful *Smart Parking Lot* (*SPL*) service process mentioned in the above use case. For the sake of simplification, we just illustrate the *searching and path planning* of the SPL. First, the end user should analyze the key data entities required by the SPL following the steps in the SCM, and enumerates six data entities of SPL. Second, the user needs to find the applicable services for the SPL based on the above data entities, which consist of five necessary services for the SPL: *locating agent*, *parking lot searching*, *client input agent*, *route planning*, and *client output agent*. Third, the end user constructs the dependency relationships among the services. Finally, the user should drag and drop the data flow graph of SPL in the SCE. The illustration of the development of *SPL* with the SCM is shown in Fig. 5.

Lessons Learned

The first lesson is that the introduction of a lightweight service mashup platform provides the capabilities for ordinary users to use and invoke telecom services into their existing web pages easily without any programming requirements. Unlike the existing open APIs and graphical SCEs of telecom networks, which are designed for experienced developers and are difficult for ordinary users to use, this platform can support ordinary Web 2.0 users in utilizing telecommunications service composition in an ad hoc fashion and substantially lowers the technical threshold. It is possible to cultivate a potentially promising long-tailed market for telecom operators.

The second lesson is that the granularity of the basic services used in service mashup is an important parameter related to the intuitiveness of service composition. In other words, higher

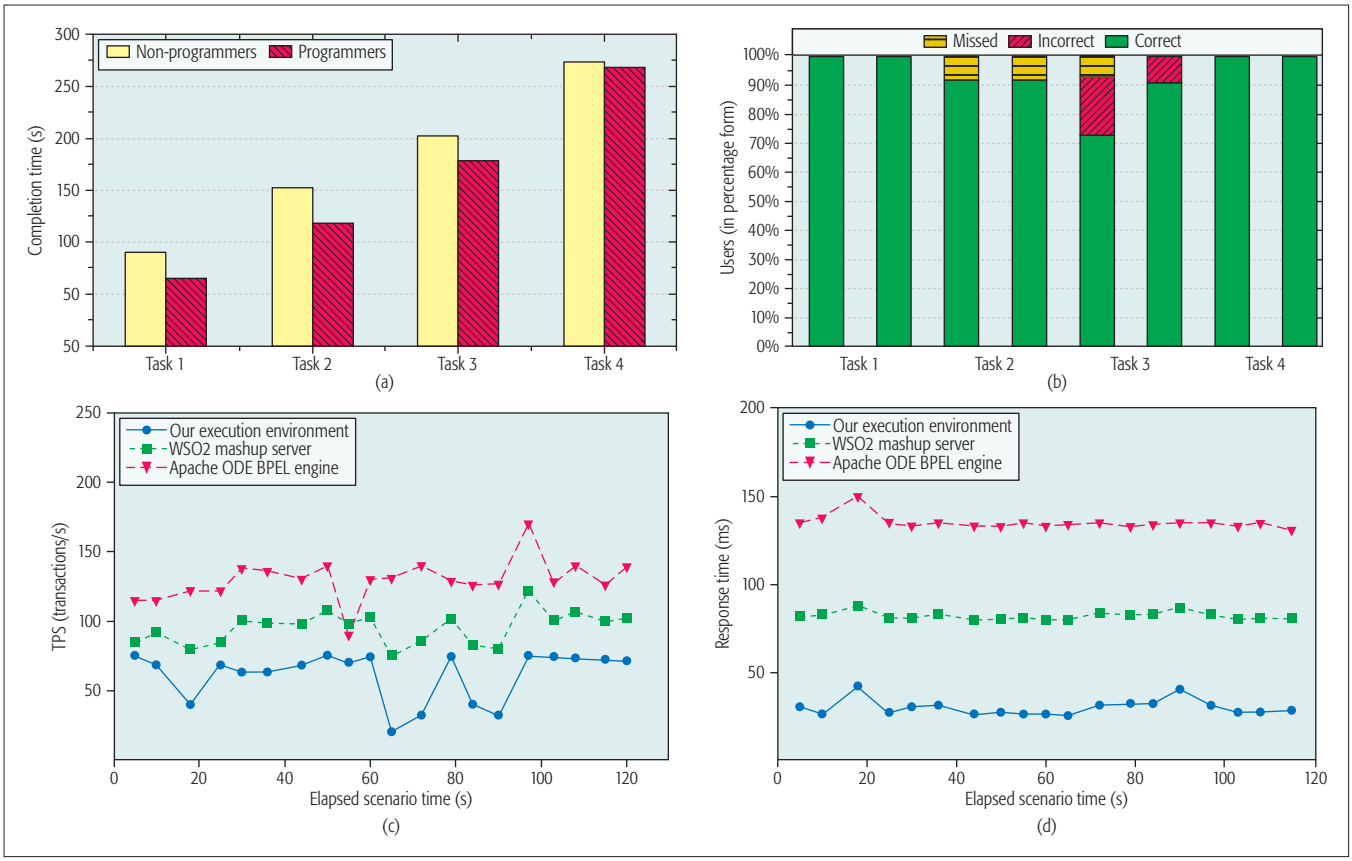


Figure 6. Experiments: a) completion time; b) accuracy and efficiency; c) comparison of transactions per second; d) comparison of response time.

abstraction levels result in more intuitive SCEs. To help users understand the features provided by the available services and the effect that each service may have on the overall composition, we must develop representations of services as visual objects that abstract from technical details, for example, their programmatic interface or communication protocol. Users should be asked to manipulate (e.g., add, remove, or modify) visual objects by operating service visualization properties rather than required to configure the technical details of services and the composition logic.

The third lesson relates to service composition support. Ordinary users do not have extensive development knowledge. Providing the graphic editor and interface could shield the ordinary user from one of the most complex aspects of mashups (i.e., data mappings). Users only need to think about the data flow, and the components should know which data to use. We also observed that ordinary users lack algorithmic thinking abilities and would benefit from receiving not only suggestions for individual event/task descriptions but also hints regarding how to couple them in the right order.

The fourth lesson is that continuous feedback is required for service execution. To further enhance users' perceptions of the effects that individual actions or services have on the final applications and allow users to understand the current state and feel of a composition, it is highly desirable to provide immediate visual feedback on any composition action and support the immediate execution of the resulting mashup. This requirement is supported by our observations indicating that end users typically have difficulty understanding the difference between design time and runtime.

Experiments

Usability Experiment for SCE

We recruited a total of 22 participants from young university staff and student volunteers from the Beijing University of Posts and Telecommunications. Participants were asked to fill out a questionnaire regarding their computing and research evaluation skills before the test, to watch a video tutorial about the LSMP platform, and to use the tool. This interaction was filmed, as was the interview that followed task completion. In terms of programming skills, they were equally divided into programmers and non-programmers. The participants were given four tasks with different complexity, which means the number of APIs to be mashed up increased, after receiving a short tutorial for five minutes. The result of accuracy and efficiency is shown in Figs. 6a and 6b: most participants completed all the tasks correctly within three minutes. Out of the 88 total tasks, the 11 programmers completed 86 correct tasks, and the 11 non-programmers completed 84 correct tasks. Moreover, the final level of satisfaction result is reported for the participants: about 72 percent of participants felt satisfied with the LSMP, 89 percent were interested in continuing to use LSMP tool, and 87 percent wanted to recommend it to their app development partners.

Response Time Experiment for SEE

We also compared the response time for our proposed SEE and those of the WSO2 mashup server and a conventional request-response-based services coordination engine. In this test, the Apache Orchestration Director Engine (ODE) was used as the centralized web service orchestration and coordination engine. BPEL was used to orchestrate a request-response-based *SPL* service. Figure 6c shows the number of transactions per second (TPS) for different approaches. For the conventional request-response services coordination approach, the maximum number of transactions per second is 74, and the average number of transactions per second is 64. For the WSO2 mashup server, the maximum number of transactions per second is 118, and the average number of transactions per second is 88. However, in the case of our LSMP, the maximum number of transactions per second is 167, and the average number of transactions per second is 130. Figure 6d shows the corresponding response times for the three different approaches. For the Apache ODE BPEL engine, the maximum response time is 150 ms, and the average response time is 134 ms; for the WSO2 mashup server, the maximum response time is 90 ms, and the average response time is 78 ms; for our LSMP, the maximum response time is 42 ms, and the average response time is 30 ms. Thus, the processing speed of LSMP is faster than those of the Apache ODE BPEL engine and WSO2 mashup server. This is because in those systems, there is a central control point that handles the "request-response" BPEL process, which may cause all messages to propagate along a longer route and wait in a central queue for processing. In contrast, our LSMP uses the asynchronous communication mechanism to support the concurrent coordinated execution of multiple services.

Conclusions

This article presents a lightweight service creation theory including the SDM, SRM, SCM, and SEM. Based on the SDM, SRM, and SCM, we implemented a lightweight SCE with web technology to provide a unified development interface to package atom-services with existing web-based services on the Internet. We also introduce an SCM to guide ordinary users in the design and implementation of new services via a data-centric approach, providing a design method for the service process with key data as the driving characteristic. We also illustrate an SPL in LSMP, and measure and analyze its performance. The results show that the lightweight service mashup platform worked well, as expected.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (no. 61132001) and the National Grand Fundamental Research 973 Program of China under grant no. 2013CB329102.

References

[1] L. Fallon and D. O'Sullivan, "The Aesop Approach for Semantic-Based Ordinary-User Service Optimization," *IEEE Trans. Network Service Management*, vol. 11, no. 2, 2014, pp. 220–34.

[2] W. Chou, L. Li, and F. Liu, "Web Services for Communication over IP," *IEEE Commun. Mag.*, vol.46, no.3, 2008, pp.136–43.

[3] F. Belqasmi, R. Glitho, and C. Fu, "RESTful Web Services for Service Provisioning in Next-Generation Networks: A Survey," *IEEE Commun. Mag.*, vol. 49, no. 12, Dec. 2011, pp. 66–73.

[4] QX. Qiao *et al.*, "Opening Up Telecom Networks with a Lightweight Web Element Service Cloud for Ordinary Users in the Web 2.0 Era," *IEEE Commun. Mag.*, vol. 52, no. 10, Oct. 2014, pp. 127–33.

[5] Y. Jung *et al.*, "Employing Collective Intelligence for User Driven Service Creation," *IEEE Commun. Mag.*, vol. 49, no. 1, Jan. 2011, pp. 76–83.

[6] T. Moriya and J. Akahani, "Application Programming Gap between Telecommunication and Internet," *IEEE Commun. Mag.*, vol. 48, no. 8, Aug. 2010, pp. 96–102.

[7] OASIS, "Web Services Business Process Execution Language (version 2.0)," http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.pdf, April 2014.

[8] N. Laga *et al.*, "Widgets and Composition Mechanism for Service Creation by Ordinary Users," *IEEE Commun. Mag.*, vol. 50, no. 3, Mar. 2012, pp. 52–60.

[9] Y. Jung *et al.*, "Employing Collective Intelligence for User Driven Service Creation," *IEEE Commun. Mag.*, vol. 49, no. 1, Jan. 2011, pp. 76–83.

[10] Z. Zhai *et al.*, "Design and Implementation: The End User Development Ecosystem for Cross-Platform Mobile Applications," *Proc. 25th World Wide Web Conf.*, Montreal, Canada, April 11–15, 2016.

Biographies

Bo Cheng received his Ph.D. degree in computer science at the University of Electronics Science and Technology of China in 2006. His research interests include multimedia communications and services computing. Currently, he is a professor with the State Key Laboratory of Networking and Switching Technology at the Beijing University of Posts and Telecommunications.

Zhongyi Zhai is a Ph.D. candidate with the State Key Laboratory of Networking and Switching, Beijing University of Posts and Telecommunications. His current research interests include web service composition, communication web service, and the Internet of Things.

Shuai Zhao is a postdoctoral fellow with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. His current research interests include the Internet of Things and service computing.

Junliang Chen is a professor at the Beijing University of Posts and Telecommunications. His research interests are in the area of service creation technology. He was elected as a member of the Chinese Academy of Science in 1991 and a member of the Chinese Academy of Engineering in 1994.

To further enhance users' perceptions of the effects that individual actions or services have on the final applications and allow users to understand the current state and feel of a composition, it is highly desirable to provide immediate visual feedback on any composition action and support the immediate execution of the resulting mashup.

On the Split-TCP Performance over Real 4G LTE and 3G Wireless Networks

Bong Ho Kim and Doru Calin

The authors share a sample of TCP throughput performance measurements after implementing Split-TCP on commercial 4G LTE and 3G networks. Experimental results show that the Split-TCP provides average TCP throughput gains on the order of 60 percent over a 4G LTE network and on the order of 5 percent over a 3G network.

ABSTRACT

The global mobile traffic is growing exponentially, and currently more than 90 percent of the internet traffic depends on TCP for reliable transmission. However, it is well known that TCP performs quite poorly over unreliable wireless networks, while its dynamic TCP flow control is sensitive to congestion events and tends to underutilize the available network capacity. Despite these known limitations, the industry believes that significant network performance enhancement can be achieved through TCP optimization, and in particular through solutions centered on Split-TCP. In this article we share a sample of TCP throughput performance measurements after implementing Split-TCP on commercial 4G LTE and 3G networks. Experimental results show that Split-TCP provides average TCP throughput gains on the order of 60 percent over a 4G LTE network and on the order of 5 percent over a 3G network. Furthermore, TCP throughput gains superior to 200 percent were measured for individual TCP connections over the 4G LTE network. We expect such throughput gains to be even greater when advanced radio technologies, such as 5G, are deployed. For these reasons, Split-TCP has the potential to be widely deployed in both current and next generation networks.

INTRODUCTION

Many studies estimate exponential traffic growth in the coming years. For instance, [1] forecasts that the global mobile traffic growth will increase eightfold between 2015 and 2020, and the monthly mobile data traffic will reach 30.6 exabytes by 2020; it also points out that traffic from wireless and mobile devices will account for two-thirds of total IP traffic by 2020. Hence, the demands for supporting the dynamically evolving mobile applications and improving network performance become increasingly critical for the operators. Particularly, the performance of TCP is critical to the end-to-end application performance, since more than a decade of internet traffic analysis consistently shows that more than 90 percent of Internet traffic uses TCP in the network [2, 3]. As a higher layer protocol, TCP traffic flows across both wired and wireless networks; thus, as the wireless traffic volume grows exponentially, the optimization of TCP performance over wireless networks is of critical importance.

Because of the wide range of usage and popularity of the current Internet protocol (TCP/IP), it is very likely that it will be used even more in the future, as the Internet Engineering Task Force (IETF) standardization community continues to improve TCP. For example, one of the activities in the TCP research community for the past several years has been targeting a multipath TCP (MP-TCP) solution [4] that establishes multiple parallel TCP connections with multiple addresses instead of a single TCP connection. This requires the TCP protocol stack to be modified, but it still runs over traditional TCP, since each of the TCP sub-flows is operated as a traditional TCP connection. Thus, the TCP performance improvement with a single TCP connection is still critical for end-to-end application performance enhancement.

Although TCP is such a popular protocol, it is also known to perform poorly over challenging radio conditions, where packet losses due to transmission errors are misintepreted as indications of network congestion. One of the mechanisms that may significantly improve TCP performance under such challenging radio environments is "Split-TCP." However, the adaptive nature of Split-TCP, which still relies on the dynamic TCP flow control in response to fluctuating network conditions, remains to be well understood and well quantified under realistic operational network conditions. One of the open points, for instance, is related to quantifying the perceived performance gain by the end user, which could be attributed to the Split-TCP mechanism. To answer some of these questions, we deployed a Split-TCP system in a commercial fourth generation (4G) LTE and 3G operator's network in multiple locations.

In this article, we present a TCP packet trace analysis; measurements collected from seven independent geographical locations allow us to evaluate the TCP performance gains that are attributed to Split-TCP. The rest of the article is organized as follows. The next section describes the fundamental LTE packet transmission procedures and the Split-TCP mechanism. The network environment along with the measurement scenarios for the TCP performance analysis are described in the following section. We then show the measurement and analysis results. Finally, the last two sections are devoted to lessons learned and concluding remarks.

Digital Object Identifier:
10.1109/MCOM.2017.1600369

The authors are with Nokia Bell Labs.

LTE Packet Transmission Procedures and Split-TCP

The purpose of TCP is to provide reliable, connection-oriented service to the application layer. Therefore, its end-to-end reliability between a host and a data network is an important design criterion. TCP was originally designed for wired links, where the packet loss rate is typically low and packet losses are due to congestion in the network. On the other hand, wireless systems are more prone to errors in the radio channel, which are caused by various impairments such as fading, shadowing, interference, and propagation losses. These effects are challenging to a transport layer protocol such as TCP, and are causes of performance degradation. The protocol is sensitive in particular to high bit error rates and losses, long end-to-end delays, interruptions in packet transmission, and wireless link rate variability. The following subsection focuses on the LTE packet transmission procedures and points out the rationale for possible residual packet losses over the LTE air link that may cause degradation in TCP performance.

LTE Packet Transmission Procedures

In LTE, retransmission of lost packets is handled by the hybrid authomatic repeat request (HARQ) mechanism in the medium access control (MAC) layer and by the ARQ mechanism in the radio link control (RLC) layer to increase the reliablity of packet delivery over the air link. LTE technology relies on these two mechanisms to fight the errors associated with packet transmissions over lossy radio channels. In particular, through its fast retransmission capability, HARQ is designed to combat dynamic changes in radio link conditions caused by fading and interference, as discussed below.

Figure 1 illustrates the LTE data flow of an IP packet from the Packet Data Convergence Protocol (PDCP) layer through the MAC layer. This figure shows the functionalities of each layer, including segmentation, concatenation, multiplexing, and adding protocol headers to the data units. PDCP receives upper layer IP packets as input data and adds PDCP headers on top of them. The resulting packets are referred as RLC service data units (SDUs). The RLC uses segmentation and concatenation mechanisms to fill up the RLC payload of the RLC packet data units (PDUs). When an RLC SDU cannot be included in a given RLC payload because the unnocupied size of the corresponding RLC PDU is too short, the SDU is segmented and transmitted using multiple PDUs. Inversly, if the SDU size is smaller than the PDU, the RLC layer concatenates multiple SDUs in order to fill up the RLC PDU payload. The RLC processes the SDUs in order to allow delivery of SDUs in the correct order.

The main purpose of the RLC protocol layer is to deliver a data packet from/to its peer RLC layer via one of the three transmission modes: transparent mode (TM), unacknowledge mode (UM), and acknowledge mode (AM). The RLC AM mode is able to retransmit packets in case a loss is detected. The ARQ retransmissions are triggered by either an RLC status report exchanged between peers or a failure of a HARQ process to deliver a transport block (TB).

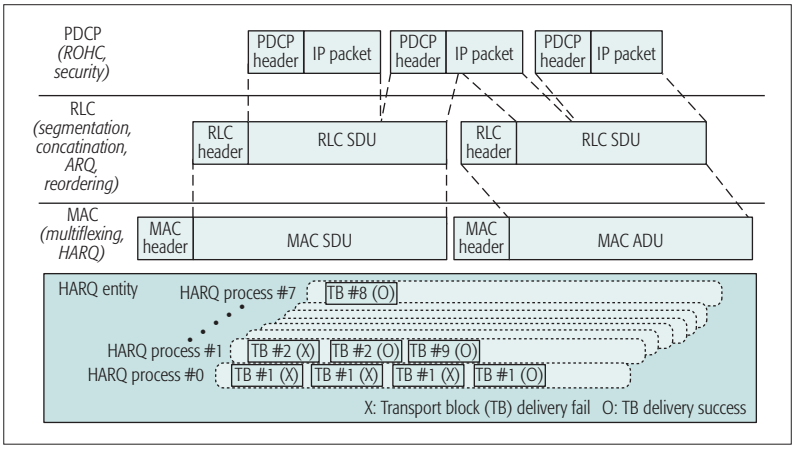


Figure 1. LTE data flow through the layer 2 protocol stack.

In real operational environments, the HARQ settings are determined based on the trade-off between the desired residual error target post HARQ transmissions and the affordable delay budget (e.g., more packet retransmissions may further reduce the packet loss rate, but at the expense of more packet delay in the HARQ processing). HARQ has the ability to buffer transport blocks that were not received correctly and to perform soft combining across the buffered data and the retransmitted transport blocks. The HARQ functionality spans the physical and MAC layers; the HARQ protocol is part of the MAC layer, while the soft combining operation is performed at the physical layer. There is one HARQ entity at the MAC layer, and it maintains a number of parallel HARQ processes. Each HARQ process is associated with a HARQ process identifier [5]. For an LTE frequency-division duplex (FDD) system, there shall be a maximum of eight parallel stop-and-wait HARQ processes in the downlink, so while some HARQ processes may be locked in packet retransmissions, transmission of fresh packets is still possible through the other available HARQ processes. The mobile terminal, referred to as user equipment (UE), determines the channel quality indicator (CQI) required for link adaptation (LA) during packet transmission based on measurements of the downlink reference signals. The CQI corresponds to the highest modulation and coding scheme (MCS) that allows a UE to operate under the assumptions of transport block error rate (BLER) not exceeding 10 percent during the first packet transmission under specific radio channel conditions [6]. BLER is a measure of how successful a data transmission is over the air at the physical and MAC layer levels. In general, the BLER decreases with each packet retransmission associated with the HARQ process, and in many practical cases only a small number of retransmissions (e.g., two through four) are required to bring the post HARQ packet error rate down to 1 percent. Nevertheless, such a low packet error rate is not satisfactory at the TCP layer, and for this reason, the residual packet loss after reaching the maximum HARQ retransmission attempts may be further corrected by the ARQ mechanism in the RLC layer. This additional error correction processing in the RLC layer is expected to bring

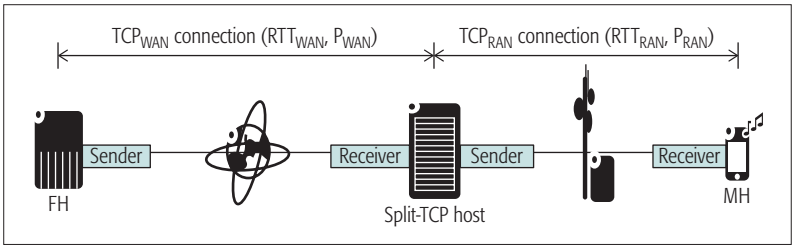


Figure 2. Conceptual diagram of Split-TCP.

down the residual packet loss rate to levels as low as 10^{-6}, which is consistent with the Third Generation Partnership Project (3GPP) performance guidelines [7] in order to maximize the TCP layer performance.

Referemce [8] provides an LTE measurement study on throughput, packet delay, and packet loss in an LTE network. This study examines the uplink traffic for the transmission rate of 50 Mb/s. One observes packet losses throughout the entire trace and finds out that most packet losses are bursty. When packet losses occur, the affected burst length is frequently on the order of 20 packets and can be even larger than 100 packets. This may be attributed to the nature of the wireless channel and to its temporal correlations [8].

Split-TCP Solution

As aforementioned, the TCP throughput is signifcantly affected by the source-destination round-trip time (RTT) and packet loss rate. Since the TCP requires receiver acknowledgments for every window of data packets sent, TCP throughput is inversely proportional to the end-to-end RTT (RTT_{e2e}) and the square root of the end-to-end packet loss rate (P_{e2e}), as shown in Eq. 1 in [9]. Thus, the distance and the packet loss rate between the server and the end user become bottleneck factors for the TCP throughput and are hard to overcome, unless the server is relatively close to the end user.

The Split-TCP is one of the mechanisms that may significantly improve the TCP throughput performance under challenging network conditions, by segmenting the RTT_{e2e} across different network segments and by isolating the packet loss events to those network segments that are the root cause for problems if there are multiple network segments which have different network characteristics. The conceptual Split-TCP diagram is illustrated in Fig. 2, where the TCP_{RAN} is the connection between a mobile host (MH) and a Split-TCP host (SH), and the TCP_{WAN} is the connection between a far host (FH) and the SH.

Split-TCP divides an end-to-end TCP_{e2e} connection characterized by the delay RTT_{e2e} and packet loss rate P_{e2e} into two independent TCP connections: a TCP connection in the radio access network (TCP_{RAN}) with the RTT delay (RTT_{RAN}) and packet loss rate (P_{RAN}) for the corresponding TCP_{RAN} connection and another TCP connection in the wide area network (TCP_{WAN}) with the RTT delay (RTT_{WAN}) and packet loss rate (P_{WAN}) for the corresponding TCP_{WAN} connection, as depicted in Fig. 2 and represented in Eq. 1.

$$TP \sim \frac{1}{RTT_{e2e}*\sqrt{P_{e2e}}} \sim \frac{1}{(RTT_{RAN}+RTT_{WAN})*\sqrt{P_{WAN}+P_{RAN}}} \quad (1)$$

$$ETP \sim \min\frac{1}{(RTT_{WAN})*\sqrt{P_{WAN}}}, \frac{1}{(RTT_{RAN})*\sqrt{P_{RAN}}} \quad (2)$$

This has the effect of isolating the packet loss to the network segment that is responsible for the packet loss occurrence. Hence, the effective end-to-end TCP throughput is the minimum throughput across the individual TCP connections resulting from splitting the link, and the improvement is expressed in Eq. 2. One should remark, however, that the TCP throughput gain depends on the split point in the network, which is defined by the RTT_{e2e} split ratio and by the packet loss rates in the TCP_{RAN} and TCP_{WAN} segments.

There are various types of Split-TCP mechanisms, including Indirect-TCP (I-TCP) [10], Aggregate TCP (ATCP) [11], TCP for Mobile Cellular Networks (M-TCP) [12], M-TCP+ [13], Enhanced Split-TCP (ES-TCP) [14] and Advanced Split-TCP (AS-TCP)[15]. The I-TCP solution has capabilities for both packet loss event isolation and RTT split and provides faster recovery due to relatively shorter RTT on wireless links, but it does not maintain the end-to-end protocol semantics. M-TCP and M-TCP+ are the options that maintain the end-to-end protocol semantics, but they aim to improve the TCP efficiency, while the wireless link suffers disconnections by preventing the sender from closing down the congestion window to the minimum. The SH using the M-TCP and M-TCP+ mechanisms freezes the packet transmission from the FH when an air link disconnection is detected. These mechanisms still propagate the packet losses on the wireless network segment into the wired network, since the acknowledgment (ACK) packets from an MH are instead forwarded to the FH. Consequently, these mechanisms lose a main benefit of Split-TCP, which is shortening the RTT delay for TCP. Shortening the RTT by itself could increase the TCP performance significantly, even without link condition changes. Among these split TCP mechanisms, ES-TCP and AS-TCP are the only solutions that have capabilities for both packet loss event isolation and RTT split while maintaining the end-to-end protocol semantics.

Network Environment and Performance Measurement Scenarios

We deployed a Split-TCP system in commercial 4G LTE and 3G networks in the North American market in order to evaluate the end-to-end TCP throughput performance and the perceived end-user experience. The network is characterized by seven independent geographical regions (i.e., R1 through R7), as listed in Table 1, and separate Split-TCP systems are deployed in each of the seven regions; this table summarizes the combination of the measurement scenarios considered for this analysis. Since the end-to-end network delay observed during file downloads depends

on the relative location of servers and clients, we have selected two extreme locations for the servers. The two servers, identified as S1 and S2, are more than 2000 mi apart from each other. As for the location of clients, we identified one cell site per region and collected measurements, while the number of connected clients is representative of a typical user in a moderately loaded cell. For example, the peak number of connected subscribers is about or below 100 during the measurement in the selected cell site. Furthermore, we have selected three client locations within the tested cell sites, which are referred to as near-cell (NC) for good radio conditions, mid-cell (MC) for average radio conditions, and far-cell (FC) for poor radio conditions. The guidelines for selecting these three locations are specified in Table 1 with respect to received signal reference power (RSRP) zone thresholds. The motivation for these placements is to account for the fact that the maximum available bandwidth depends on the client's radio conditions and to get a fair sampling of the perceived performance within a cell. Furthermore, we employed five different file sizes through these experiments: 0.5 MB, 1 MB, 5 MB, 10 MB, and 20 MB. With these conditions, we collected TCP packet traces at the network interface in the Split-TCP host (SH) toward the RAN by repeating 20 independent file download experiments for each file size, with and without TCP optimization (i.e., Split-TCP) in order to allow performance comparison. All test scenarios were executed over both 3G and 4G LTE networks.

Two server locations	S1 and S2 (> 2000 miles apart)
Seven cell sites	R1, R2, R3, R4, R5, R6, R7
Three client locations per cell site	Near cell (NC): RSRP > –80 dBm Mid cell (MC): –100dBm < RSRP < –80dBm Far cell (FC): RSRP < –100 dBm
File sizes for 4G LTE	0.5 MB, 1 MB, 5 MB, 10 MB, 20 MB
File sizes for 3G	0.5 MB and 1 MB

Table 1. Measurement criteria.

Measurement Results and Analysis

We extracted various information from the TCP packet traces including download time, throughput, goodput, ratio of retransmitted data bytes, RTT measurement, duplicate ACK packet ratio, number of out-of-order packets, and amount of in-flight bytes. In this article we show the TCP goodput gain and the ratio of retransmitted data bytes. We have noticed that the TCP goodput results for the baseline (i.e., without TCP optimization) vary widely for some test sites, and in particular for the FC locations. This variability is likely caused by the fluctuation in the network condition. Since the measurements were done over an uncontrolled live network, some of the performance gain comparisons may not be sufficiently reliable with a small number of measurements.

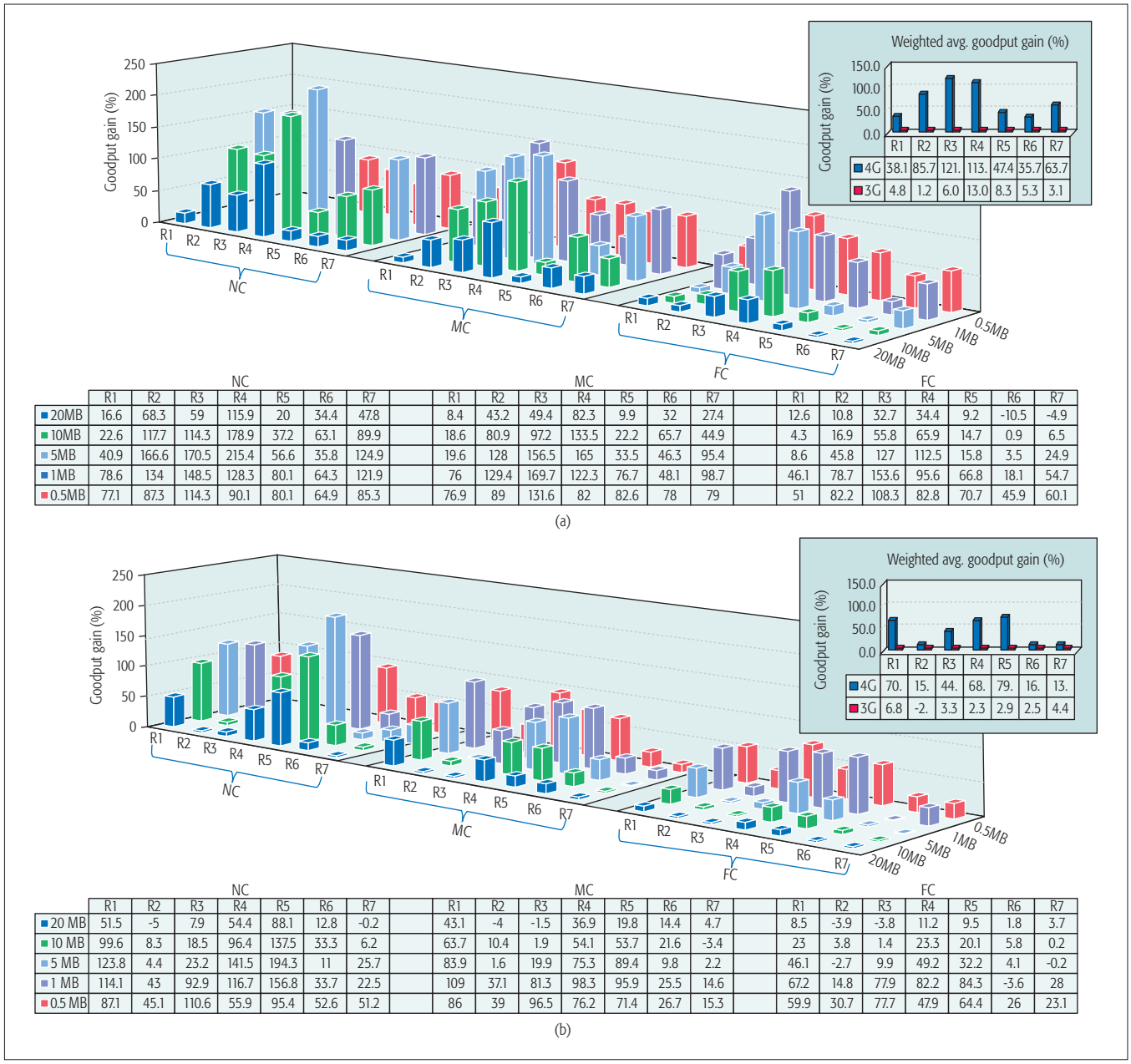
Figure 3 shows the TCP goodput performance gain over the LTE network, in contrast to the performance with and without TCP optimization. The percentage values in the associated tables are obtained from averaging across 20 repetitive measurements performed per measurement scenario according to Table 1. The seven measurement regions are indexed from R1 to R7 and the test client locations (i.e., NC, MC, and FC) are clustered together. Figure 3a shows the measurement results using server S1, and Fig. 3b shows the results using server S2. The TCP goodput gain range is widely spread, and it is as high as 215 percent, which means that the TCP optimization provides goodputs that are more than three times faster than the baseline performance (without TCP optimization). Overall, the TCP goodput gains are more modest in FC conditions. Because of the wide range of performance gains, we set two criteria to calculate an aggregated weighted TCP goodput gain per measurement region, and the weighted contribution values per RF condition are selected considering the RF condition distribution of the connected user devices in the measured cells:

- Weighted contributions per file size: 45 percent (for file sizes ≤ 1 MB), 30 percent (for 5 MB file size), and 25 percent (for file sizes ≥ 10 MB)
- Weighted contributions per RF conditions: 25 percent (for NC), 30 percent (for MC), and 45 percent (for FC)

The weighted average graphs in Fig. 3a are based on these weighted average criteria. These graphs show the average TCP goodput gains for the seven regions over both 4G LTE and 3G networks, even though the detailed measurement results over the 3G network are not shown in this article. Referring to the server S1, the region R3 shows about 120 percent of TCP goodput gain, while the overall weighted average TCP goodput gain across all the seven regions is about 72 percent over the LTE network and about 6 percent over the 3G network. Similar information is displayed in Fig. 3b for the server S2: the overall weighted average TCP goodput gain across all the seven regions is about 44 percent over the 4G LTE network and about 3 percent over the 3G network. One of the reasons for the performance gain difference between these two server locations is the relative distance between servers and clients. Five out of the seven total measurement regions are closer to server S2 compared to server S1. This is described in more detail later in this section.

Figures 4a and 4b show the percentage of retransmitted bytes per measurement scenario using (a) server S1 and (b) server S2 over the 4G LTE network. Overall, the packet retransmission ratio is noticeably low; even 0 percent packet retransmission has been observed quite often, and most of the time it is below 0.5 percent and seldom gets larger than 1 percent. These results are collected from the end-to-end TCP connections, which span across both wireless and wireline networks.

Figure 4c represents the TCP sequence number graphs obtained from all individual TCP traces that are collected from the NC locations (good radio conditions) in one of the test regions while downloading a 10 MB file from the server S1 over the 4G LTE network. The group of TCP sequence graphs in the upper part of Fig. 4c is collected



	NC							MC							FC						
	R1	R2	R3	R4	R5	R6	R7	R1	R2	R3	R4	R5	R6	R7	R1	R2	R3	R4	R5	R6	R7
20MB	16.6	68.3	59	115.9	20	34.4	47.8	8.4	43.2	49.4	82.3	9.9	32	27.4	12.6	10.8	32.7	34.4	9.2	-10.5	-4.9
10MB	22.6	117.7	114.3	178.9	37.2	63.1	89.9	18.6	80.9	97.2	133.5	22.2	65.7	44.9	4.3	16.9	55.8	65.9	14.7	0.9	6.5
5MB	40.9	166.6	170.5	215.4	56.6	35.8	124.9	19.6	128	156.5	165	33.5	46.3	95.4	8.6	45.8	127	112.5	15.8	3.5	24.9
1MB	78.6	134	148.5	128.3	80.1	64.3	121.9	76	129.4	169.7	122.3	76.7	48.1	98.7	46.1	78.7	153.6	95.6	66.8	18.1	54.7
0.5MB	77.1	87.3	114.3	90.1	80.1	64.9	85.3	76.9	89	131.6	82	82.6	78	79	51	82.2	108.3	82.8	70.7	45.9	60.1

(a)

	NC							MC							FC						
	R1	R2	R3	R4	R5	R6	R7	R1	R2	R3	R4	R5	R6	R7	R1	R2	R3	R4	R5	R6	R7
20 MB	51.5	-5	7.9	54.4	88.1	12.8	-0.2	43.1	-4	-1.5	36.9	19.8	14.4	4.7	8.5	-3.9	-3.8	11.2	9.5	1.8	3.7
10 MB	99.6	8.3	18.5	96.4	137.5	33.3	6.2	63.7	10.4	1.9	54.1	53.7	21.6	-3.4	23	3.8	1.4	23.3	20.1	5.8	0.2
5 MB	123.8	4.4	23.2	141.5	194.3	11	25.7	83.9	1.6	19.9	75.3	89.4	9.8	2.2	46.1	-2.7	9.9	49.2	32.2	4.1	-0.2
1 MB	114.1	43	92.9	116.7	156.8	33.7	22.5	109	37.1	81.3	98.3	95.9	25.5	14.6	67.2	14.8	77.9	82.2	84.3	-3.6	28
0.5 MB	87.1	45.1	110.6	55.9	95.4	52.6	51.2	86	39	96.5	76.2	71.4	26.7	15.3	59.9	30.7	77.7	47.9	64.4	26	23.1

(b)

Figure 3. 4G LTE TCP goodput gain (percent) comparison: a) 4G LTE TCP goodput gain (percent): TCP_Opt_OFF vs. TCP_Opt_ON (server S1); b) 4G LTE TCP goodput gain (percent): TCP_Opt_OFF vs. TCP_Opt_ON (server S2).

without TCP optimization, while the group of TCP sequence graphs in the lower part of Fig. 4c is collected with TCP optimization. This indicates that the retransmission rate is zero percent; otherwise, there would be red dot(s) indicating packet retransmissions and duplicate ACK packets. Furthermore, the TCP goodput gain for this scenario is around 200 percent, which means that the TCP optimization renders the download speeds three times faster over the baseline performance.

On the other hand, the TCP traces in Fig. 4d corresponds to a scenario with about 0.8 percent of TCP packet retransmission caused by packet losses; in addition, about 10 percent of ACK packets are duplicate ACKs, and are illustrated as red dots. The TCP goodput gain for this scenario is about 33 percent, which is noticeably lower compared to the measurements shown in Fig. 4c. One of the reasons for the lower performance gain is the frequent retransmission timeout (RTO) expiration caused by the burst of packet loss.

Figure 5 illustrates the distinctive performance gain differences between server S1 and server S2. These measurements have been collected from one of the regions that is very close to server S2. Noticeably, the measurements taken with reference to server S1 show high performance gains, while the measurements with reference to server S2 show smaller gains. This is because the TCP performance gains depend on the available bandwidth in the network and network delays. If the network bandwidth is underutilized, higher TCP optimization gains are expected; on the other hand, if the network bandwidth is already saturated, there is not much room for gains through TCP optimization.

Before taking TCP performance measurements in Fig. 5, we tested the end-to-end network delays

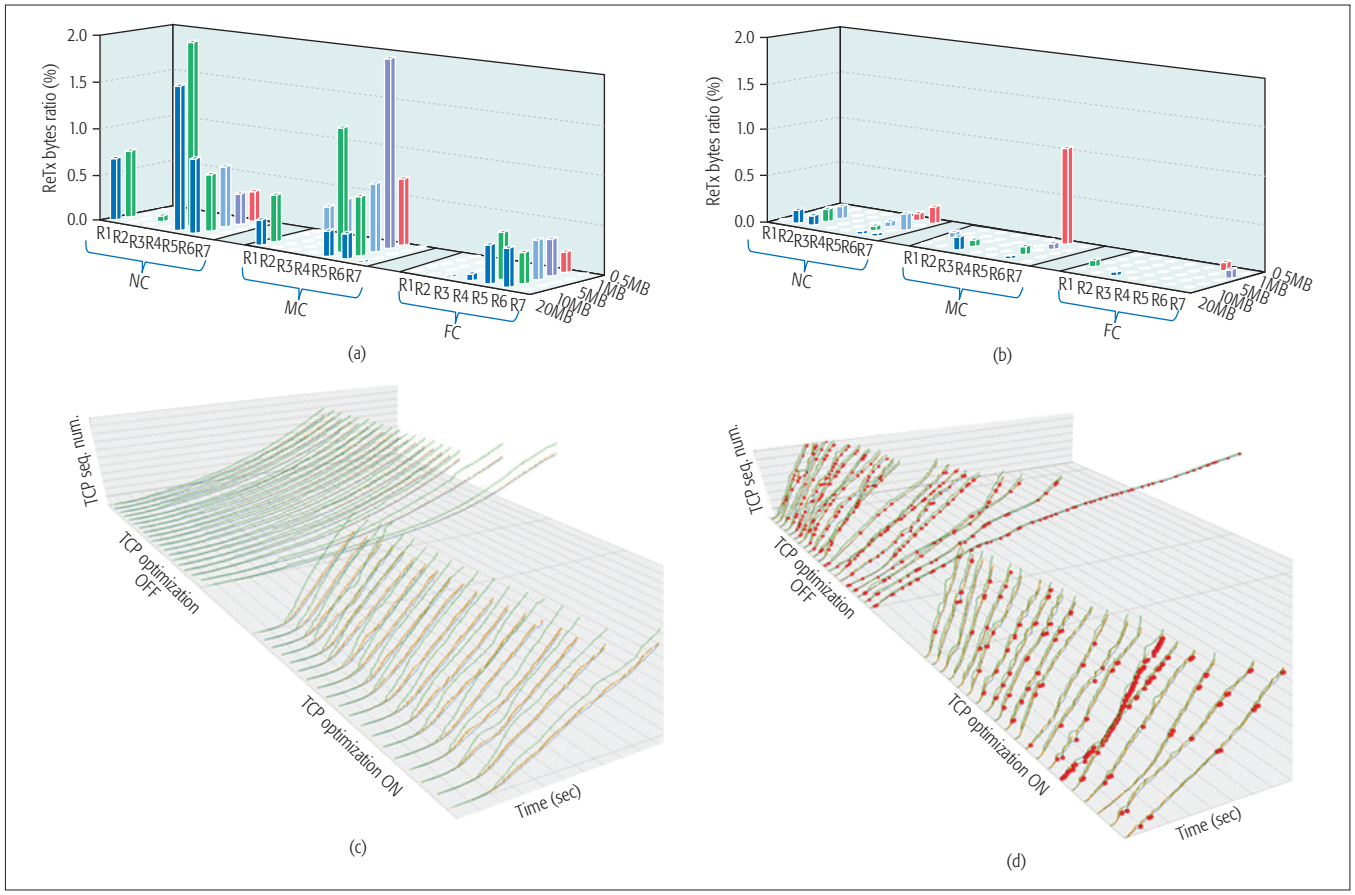


Figure 4. Retransmission bytes ratio (percent) over the LTE network: a) server S1; b) server S2, and TCP sequence number graph with: c) 0 percent retransmission and no duplicate ACKs; d) 0.8 percent retransmission and 10 percent duplicate ACKs.

and the maximum available bandwidth using speed tests taken in the proximity of the clients, presumably for measuring the available air link bandwidth. The average PING delay from a client to server S2 is about 40 ms and to server S1 is about 130 ms on average. The three speed test results identify the approximate bandwidth limits perceived at the locations labeled as NC (i.e., 25 Mb/s), MC (i.e., 20 Mb/s), and FC (i.e., 14 Mb/s). Hence, these measurements point out to the reference upper bounds per test site, and are marked as horizontal floor lines in Fig. 5a. The speed test tool gauges how much inbound traffic can be handled consistently through a connection, determining its maximum sustainable throughput (MST). In other words, it provides a reference to benchmark how mobile carriers perform at specific test locations. If the TCP throughput for the baseline configuration (without optimization) is lower than the reference benchmark, one gets a clear indication that the link is underutilized, and hence there is room for TCP throughput gains, as enabled through TCP optimization. Referring to Fig. 5b, a 10 MB file size download at an NC location results in an approximately 117 percent throughput gain resulting from TCP optimization with respect to file downloads from server S1 (which is further away from the client), while the gains are capped to 19 percent with respect to file downloads from server S2 (which is closer to the client). This is because for the baseline configuration (in the absence of TCP optimization), the actual TCP throughput measured through file downloads from server S1 was limited to 9 Mb/s, which is significantly lower than 25 Mb/s measured via speed tests, while the TCP throughput measured through file downloads from server S2 was 21 Mb/s, which is much closer to the 25 Mb/s benchmark. Post TCP optimization, the TCP throughput with server S1 increased from 9 Mb/s to 21 Mb/s, which translated into the 117 percent gain since the reference baseline for server S1 was significantly lower in comparison to server S2.

Lessons Learned and Challenges

The poor TCP performance over wireless networks is a well-known issue, and it is critical to reduce the server-client RTT, to isolate packet errors within the wireless network, and to take effective corrective actions in order to improve the TCP throughput performance. The TCP performance analysis in this article has clearly shown the TCP performance limitations over commercially deployed wireless networks, as well as the significant performance improvement achieved through the Split-TCP optimization. In addition to simply splitting a TCP connection, various TCP congestion control mechanisms and TCP optimization parameters have a major impact on the TCP performance. Hence, another important lesson is that the TCP optimization parameters should be carefully selected and adjusted dynamically as much as possible. This is a very challenging task, since each TCP connection may require different optimization configurations, depending on its own condition (e.g., based on radio con-

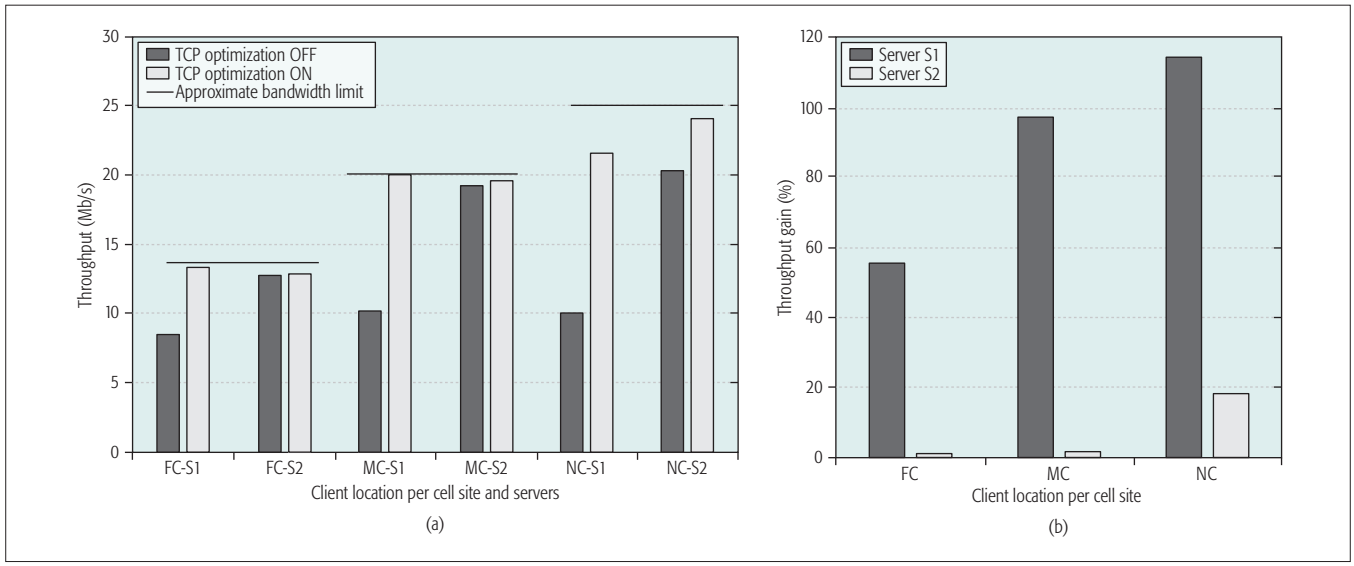


Figure 5. Throughput and gain comparison: a) throughput comparison for server S1 vs. S2 per test site; b) throughput gain comparison per cell site.

ditions, network delay, or traffic type), which is a challenging task. We should also mention that the frequent TCP optimization in real networks calls for more automation.

Conclusion

The Split-TCP mechanism clearly shows a significant TCP throughput improvement, quantified as up to 120 percent in terms of weighted average TCP goodput gain over a commercial 4G LTE network. On the other hand, 3G networks operate at lower data rates compared to 4G LTE networks and show rather modest gains through TCP optimization, limited to 6 percent by the same metric of weighted average TCP goodput gain. One should emphasize that because of the significant differences between the achievable air interface rates over 3G and 4G networks, 3G networks saturate much faster compared to 4G networks when operating under similar network conditions characterized by network delay, packet loss rate, and network congestion levels. Because of this, the room for performance gains through Split-TCP is limited in 3G networks. On the other hand, the 4G operational capacity has a broader operational range and present more opportunities for performance improvement via Split-TCP, aspects that are well represented in Fig. 3. Furthermore. to evaluate the operational merits of Split-TCP-based solutions, we have considered two server locations with respect to reference clients during file download operations: a far distance server and a much closer one. The actual network RTT delay difference between a given client and these two server locations depends on the explicit location of the client at the time of measurments; for reference, we have observed an RTT delay difference with respect to the two servers, which was as large as 90 ms. Regardless, the overall throughput gains enabled via TCP optimization for both server locations were about 60 percent on average for the 4G LTE network and about 4 percent on average for the 3G network. Significantly larger throughput gains (e.g., in excess of 200 percent), which are attributed to the TCP optimization, were noticed through individual measurements. These field experiments indicate that the performance gains enabled through a Split-TCP optimization may be maximized when the end-to-end RTT between a client and a server is large and when the network bandwidth is not saturated. Thus, we expect that the throughput gains enabled via Split-TCP optimization may be even greater when advanced radio technologies, such as 5G, are deployed since 5G technologies will unleash much more capacity compared to 4G LTE, and the room for performance improvement will be much broader. The implication of such significant TCP throughput gains is a better quality of experience for the end users, which can translate into metrics such as increased data transmission rate and reduced number of video stalling events. For these reasons, Split-TCP has the potential to be widely deployed in both current and next generation mobile networks.

References

[1] Cisco, "Visual Networking Index: Forecast and Methodology 2015–2020," ; http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html, accessed Jan. 2017.

[2] M. Fomenkov *et al.*, "Longitudinal Study of Internet Traffic in 1998-2003," *Proc. Winter Int'l. Synp. Info. Commun. Technologies*, Jan. 2004, pp. 1–6.

[3] CAIDA, "Analyzing UDP usage in Internet traffic,"; https://www.caida.org/research/traffic-analysis/tcpudpratio/, accessed July 2015

[4] A. Ford *et al.*, "RFC 6824: TCP Extensions for Multipath Operation with Multiple Addresses," Jan. 2013.

[5] 3GPP TS 36.321: "E-UTRA MAC Protocol specification," v. 13.4.0, Dec. 2016.

[6] 3GPP TR 36.213: "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures," v. 13.4.0, Dec. 2016.

[7] ETSI TR 136 912: "LTE; Feasibility Study for Further Advancements for E-UTRA (LTE-Advanced) ," v. V13.0.0, Jan. 2016

[8] N. Becker, A. Rizk, and M. Fidler, "A Measurement Study on the Application-Level Performance of LTE," *Proc. IFIP Networking*, June 2014, pp 1–9.

[9] T. Ott, J. Kemperman, and M. Mathis, "The Stationary Behavior of Ideal TCP Congestion Avoidance," 1996; http://www.teunisott.com/Papers/TCP_Paradigm/TCPwindow.pdf, accessed Jan. 2016

[10] A. Bakre and B. R. Badrinath, "I-TCP: Indirect TCP for Mobile Hosts," *Proc. 15th IEEE Int'l. Conf. Distrib. Comp. Systems*, Vancouver, BC, Canada, May 1995, pp 136–43.

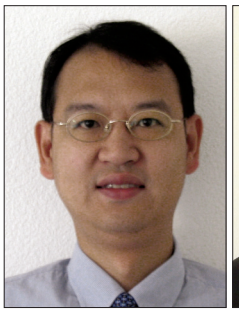
[11] R Chakravorty *et al.*, "Flow Aggregation for Enhanced TCP over Wide-Area Wireless," *Proc. IEEE INFOCOM*, Apr. 2003, pp. 1754–64.
[12] K. Brown and S. Singh, "M-TCP: TCP for Mobile Cellular Networks," *Comp. Commun. Rev., ACM SIGCOMM*, vol. 27, no. 5, Oct. 1997, pp. 19–43.
[13] M. Omueti and L. Trajkovic, "M-TCP+: Using Disconnection Feedback to Improve Performance of TCP in Wired/Wireless Networks," *Proc. SPECTS*, San Diego, CA, July 2007, pp. 443–50.
[14] B. Kim, D. Calin, and I. Lee, "Enhanced Split TCP with End-to-End Protocol Semantics over Wireless Network," to be published, *Proc. IEEE WCNC*, San Francisco CA, Mar. 2017.
[15] B. Kim, D. Calin, and I. Lee, "Advanced Split TCP with End-to-End Protocol Semantics over LTE Wireless Network," *Proc. IEEE GLOBECOM*, Washington DC, Dec. 2016, pp. 1–7.

Biographies

Bong Ho Kim [M] (bongho.kim@nokia-bell-labs.com) is a member of technical staff in Bell Labs' End-to-End Mobile Networks Solutions group in Murray Hill, New Jersey. His current research includes next generation broadband wireless technologies, cross protocol layer performance optimization, multi-radio access technology, and SDN/NFV. He has received a number of awards including the Proud Eagle Award from Lucent Technologies and the WiMAX Forum individual contribution award. He received a B.S. with honors in computer and information science from The Ohio State University, and received both M.S. and Ph.D. in computer science and engineering from the University of Pennsylvania.

Doru Calin [S'95, M'99, SM'05] (doru.calin@nokia.com) is a Bell Labs Fellow, Director, and Domain Leader with Nokia Mobile Networks CTO in Murray Hill, New Jersey, with broad responsibilities for accelerating innovations in 5G, mobile network virtualization, mobile edge cloud computing, cloud-based technologies, IoT, and verticals. He serves also as an adjunct professor with Columbia University, New York City, and as an Associate Editor of *IEEE Communications Letters*. At Nokia he held a variety of positions in research, applied research, and management of research. He was a technical consultant with Bouygues Telecom and a senior research engineer with Motorola Research Labs, Paris. He received a Ph.D. (Hons.) degree in electrical and computer engineering from the University of Versailles and Telecom SudParis, France in 1998.

INTEGRATED CIRCUITS FOR COMMUNICATIONS


Charles Chien


Zhiwei Xu

In this issue of the Integrated Circuits for Communications Series, we have selected two articles that highlight recent progress in the integrated circuits design of communication systems geared to support the rapid increase in demand for capacity. With more than one billion smartphones used worldwide, operators are scrambling to provide the most attractive package for data services that could capture new users to their networks. However, scarce spectrum limits the increased use of the network for high bandwidth video streaming, videoconferencing, and gaming applications.

Cellular networks primarily utilize bands allocated throughout the L-band, as well as the sub-1 GHz and sub-6 GHz ranges. Almost none of these bands are contiguous, and each occupies only a small portion of the total available aggregated bandwidth. Traditionally, at any given time, a cell phone would transmit/receive within one of these allocated bands (e.g. Band 7, which occupies 140 MHz bandwidth). However, such spectral allocation would be inadequate to support a large number of subscribers using high data rate services. One way out of this predicament is to aggregate bandwidths from existing band allocations to increase throughput. To this end, in the last few releases of the Third Generation Partnership Project (3GPP) standard, inter-band carrier aggregation (CA) has been defined to allow cellular phones to transmit and receive concurrently over non-contiguously allocated bands worldwide. For bands with larger bandwidths (e.g., 100 MHz), intra-band CA is defined to allow aggregation of sub-bands.

While the concept of CA is straightforward, its implementation has proven to be quite challenging. In particular, CA stresses the design of the RF front-end. The reasons are twofold. First, over the last several years, the number of supported CA combinations has risen exponentially. Second, multiple transmitters are co-located with multiple receivers on the same mobile device. The proximity creates an extremely challenging interference environment. The large number of CA combinations, if not dealt with properly, can lead to an impractically complex design, while the self-interference inherent in a CA design can render the device useless.

In the article "LTE-Advanced Pro RF Front-End Implementations to Meet Emerging Carrier Aggregation and MIMO Requirements," the authors discuss circuit architecture trade-offs that need to be considered to realize RF front-ends supporting CA. A flexible switch-combined architecture is described that enables efficient reuse of multiple duplexers and switches to achieve challenging isolation requirements among the multiple CA bands while reducing complexity and cost. Based on these design considerations, the authors highlight a fully integrated RF front-end module that supports B1/B25/B3/B4/B39 (mid bands) and B30/B40A/B7/B41 (high bands) and all associated globally required CA combinations, as well as support for upcoming 4×4 multiple-input multiple-output (MIMO) designs.

The duplexer within an RF front-end module also enables frequency-division duplexing (FDD), which allows two-way communications with sufficiently low delay required for real-time applications. In FDD, two frequency sub-bands separated by the duplex spacing are allocated for simultaneous transmit and receive. This essentially drops spectrum utilization and therefore capacity by 50 percent since to support a user in the network, two separate frequency channels are needed: one for the transmitter and one for the receiver. Alternatively, time-division duplex (TDD) utilizes two time slots in a ping-pong fashion, whereby one time slot is used for transmit and the other for receive. TDD also reduces the spectral efficiency by 50 percent.

In the last several years, a concept referred to as full-duplex has received much attention. The idea is to transmit and receive simultaneously with no time or frequency separation as is done in conventional TDD or FDD systems. If realized, full-duplex can therefore improve capacity twofold. However, such an approach requires the receiver to prevent a co-located transmitter from saturating its own electronics during transmissions, which usually overpowers the received signal by more than 100 dB.

The second article, "Integrated Full-Duplex Radios," presents integrated circuit implementation using low-cost complementary metal oxide semiconductor (CMOS) to realize several cancellation approaches needed to combat the strong self-interference inherent in full-duplex radios. These include RF frequency domain adaptive filtering, feedback utilizing dual-polarized antennas, and non-magnetic circulators. The authors also highlight the full potential of this technology when co-designed with the digital processing, as well as cross-layered optimization between the physical and medium access control layers. The article also reveals several unresolved challenges, and the authors' work represents a step toward a far-reaching goal that many will be pursuing in the near future.

We would like to take this opportunity to thank all the authors as well as reviewers for their contributions to this Series. Future issues of this Series will continue to cover circuit technologies that are enabling new emerging communication systems. If you are interested in submitting a paper to this Series, please send your paper title and an abstract to either of the Series Editors for consideration.

BIOGRAPHIES

CHARLES CHIEN (charles.chien@creonexsystems.com) is the president and CTO of CreoNex Systems, which focuses on technology development for next generation communication systems. Previously, he held key roles at Conexant Systems, SST Communications, and Rockwell. He was also an assistant adjunct professor at the University of California Los Angeles. His interests focus mainly on the design of system-on-chip solutions for communication systems. He has published in various journals and conferences, and has authored a book, *Digital Radio Systems on a Chip*.

ZHIWEI XU is currently with Zhejiang University, working on cognitive radios, high-speed ADC, and mmWave ICs. He held industry positions with SST Communications, Conexant Systems, NXP, and HRL Laboratories, where he developed wireless LAN and SoC solutions for proprietary wireless multimedia systems, CMOS cellular transceivers, multimedia over cable systems, TV tuners, software defined radios, and analog VLSI. He has published in various journals and conferences, three book chapters, and 12 granted patents.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

AMATEUR DRONE SURVEILLANCE: APPLICATIONS, ARCHITECTURES, ENABLING TECHNOLOGIES, AND PUBLIC SAFETY ISSUES

BACKGROUND

The advancement in communication, networking, computation, and sensing technologies has attracted researchers, hobbyists, and investors to deploy mini-drones, officially called unmanned aerial vehicles (UAVs), due to their enormous applications. Drones have boundless viable applications as well due to their small size and capability to fly without an on-board pilot such as in agriculture, photography, surveillance, and numerous public services. The use of drones for achieving high-speed wireless communication is one of the most significant applications for next-generation communication systems (5G). Indeed, drone-based communication network offers versatile solutions to provide wireless connectivity for devices without infrastructure coverage due to e.g., severe shadowing by urban or mountainous terrain, or damage to the communications infrastructure caused by natural disasters. But its deployment poses several public safety (PS) threats to national institutions and assets such as nuclear power plants, historical sites, and government leaders' houses because of drone's ability to carry the explosive and other destructive chemicals and agents.

In order to cope with these security threats, surveillance drones (SDrs) deployment is required for surveillance, hunting, and jamming of the amateur drone (ADr). The main motivation of deploying SDrs is to keep an eye on the ADr which can lead to serious disasters in cases where no precautionary measures are taken in a timely manner. The SDr architecture should have the capability to self-configure in case of emergency situations without the help of the central ground control station (GCS). The increasing usage of SDr in surveillance of ADr presents some challenges such as robust detection, tracking, intruder localization, and jamming. The accuracy of detection is a basic requirement of the system. In general, the accurate detection is time-consuming. In fact, a precise moving object detection method makes tracking more reliable and faster, and supports correct classification, which is quite important for SDr to be successful. The existing motion detection algorithms have the problems of computational cost and lower robustness. However, because of rapidly changing extrinsic and intrinsic camera parameters such as pan, tilt, translation, rotation, and zooming, algorithms of highest accuracy are required. Moreover, the machine learning and pattern recognition algorithms are required to detect the ADr by using the characteristics of the electromagnetic waves, sound, images that can efficiently detect the ADr.

The goal of the proposed Feature Topic (FT) is to publish comprehensive original research for all readers of the Magazine regardless of their specialty. The main objective of this FT is to bring most recent advances in amateur drone surveillance network architecture and technologies. Moreover, its goal is to address the challenges related to public safety issues posed by the flying of drone in the No-fly zone. Original research papers are to be solicited in topics of interest including, but not limited to, the following themes on applications, architectures, enabling technologies, and public safety issues with amateur drone surveillance.

- Innovative P2P and FANET cross-layer architectures and protocols for FANET
- Radio resource management schemes for surveillance drone
- Power control schemes for surveillance drone
- Spectrum sharing and future spectrum requirements for surveillance drones
- Routing and MAC protocols for drones
- Cross layer protocols for drones
- Smart and time-efficient trajectory generation schemes
- Surveillance drone control parameters optimization schemes
- Pattern recognition-based amateur drone tracking
- Machine learning algorithms to discriminate amateur drone with moving object
- Sound and electromagnetic waves detection schemes
- Thermal imaging-based amateur drone detection scheme
- Coverage extension schemes for amateur drone detection
- Signal strength, inertial sensor, and cell tower-based positioning schemes
- Active and passive object tracking schemes
- Amateur drone detection and tracking using holographic radars
- Context-aware localization schemes
- Energy efficient localization of amateur drone
- Frequency band recognition technologies for amateur drone jamming
- Location detection technologies and protocols
- 3D location detection protocols
- Disaster resilient location detection protocols
- Robust location accuracy technology development
- Multi-hop and relay-based communications
- Drone classification and identification using data mining techniques
- Interdisciplinary research for catching amateur drone
- Testbeds and experimental results of surveillance drone

SUBMISSIONS

Articles should be tutorial in nature and written in a style comprehensible and accessible to readers outside the specialty of the article. Complete guidelines for prospective authors can be found at http://www.comsoc.org/commag/paper-submission-guidelines. The Guest Editors reserve the right to reject papers they unanimously deem to be either out of the scope of this Feature Topic or otherwise extremely unlikely to be accepted after a peer review process.

It is important to note that *IEEE Communications Magazine* strongly limits mathematical content, and the number of figures and tables. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Article length (introduction through conclusions, excluding figures, tables and their captions) should not exceed 4,500 words. Figures and tables should be limited to a combined total of six (6). The number of archival references is limited to fifteen (15). All references to online sources (website URLs, web-posted papers and reports) are required to show an "Accessed on" date. All articles must be submitted through the IEEE Manuscript Central site (http://mc.manuscriptcentral.com/commag-ieee) to the "January 2018 / Amateur Drone Surveillance" category by the submission deadline according to the following schedule:

IMPORTANT DATES

Manuscript Submission Deadline: May 1, 2017

Decision Notification: September 1, 2017

Final Manuscript Due Date: October 15, 2017

Publication Date: January 2018

GUEST EDITORS

Zeeshan Kaleem (Corresponding Guest Editor)
COMSATS Institute of Inform. Technol., Pakistan
Email: zeeshankaleem@gmail.com

Mubashir Husain Rehmani
COMSATS Institute of Inform. Technol., Pakistan
Email: mshrehmani@gmail.com

Ejaz Ahmed
Nat'l Institute of Standards and Technol., Gaithersburg, USA
Email: imejaz@gmail.com

Abbas Jamalipour
Univ. of Sydney, Australia
Email: a.jamalipour@ieee.org

Joel J. P. C. Rodrigues
Nat'l Inst. of Telecomm., Brazil
Email: joeljr@ieee.org

Hassna Moustafa
Intel Corp., USA
Email: hassnaa.moustafa@intel.com

Wael Guibene
Intel Labs Europe, Ireland
Email: wael.guibene@intel.com

LTE-Advanced Pro RF Front-End Implementations to Meet Emerging Carrier Aggregation and DL MIMO Requirements

David R. Pehlke and Kevin Walsh

The authors describe best practices for meeting the challenging coexistence, harmonic management, linearity, and efficiency performance related to the functional partitioning, optimized integration, and technology selection of the RFFE.

ABSTRACT

RF front-end (RFFE) architectures and implementations are developing new ways to optimize LTE-Advanced PRO (Rel 13) multi-component carrier aggregation, advanced features to increase spectral efficiency such as higher order modulation and higher order MIMO, and the concurrent operation of all of these features together. In this article, we describe best practices for meeting the challenging coexistence, harmonic management, linearity, and efficiency performance related to the functional partitioning, optimized integration, and technology selection of the RFFE. Recent trends to improve radio performance are driving specific blocks (e.g., the low noise amplifier) into the RFFE, with associated architecture changes in both primary and diversity paths. Carrier aggregation features are supported in a number of different methods with different insertion loss, isolation, and noise figure trade-offs, and here we examine benefits of a new category of highly integrated diversity receive modules to enhance receiver sensitivity across all use cases. Movement toward higher order MIMO in the DL is compounding additional RF Rx path support and requirements, and cost-effective solutions for optimum performance trade-offs require a holistic and complete RF system view of both Tx and Rx in order to address these emerging requirements.

INTRODUCTION

As the requirements of future cellular communications are being realized, there is an enormous focus on the following top priorities for user equipment (UE) radio and RF front-end (RFFE) development:

- An incredible demand for higher data rates mandates *advanced features into the UE*.
- These features, and especially their simultaneous concurrent use, are *significantly increasing handset complexity* and performance challenges.
- More robust "always on" connections to the Internet with an acceptable cell edge user experience, even in the most challenging radio environments, are required.

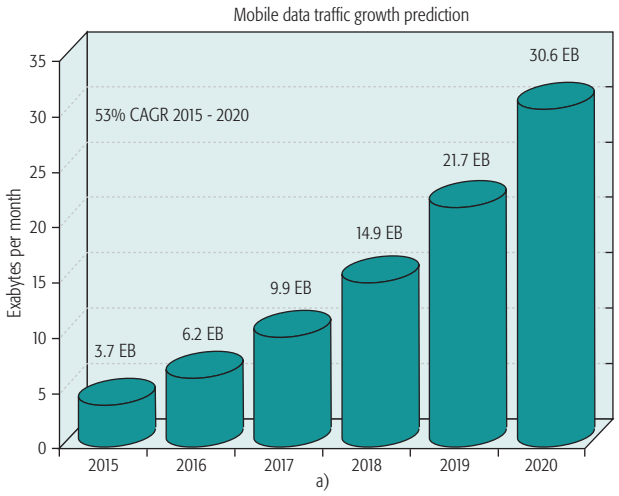
The demand for higher data rates is clear from the recently published statistics on mobile data growth [1] indicating that global mobile data traffic will grow tenfold in five years, having accelerated to reach 74 percent growth in 2015 alone. Smart devices (defined as mobile devices that have a minimum of third generation [3G] connectivity and advanced multimedia/computing capability) accounted for 90 percent of that growth figure. Mobile video traffic accounted for 55 percent of total mobile data traffic, and specifically for handsets, smartphones (including large screen phablets) were responsible for 97 percent of total global handset traffic. There is no end in sight to this overwhelming trend toward big mobile data enabled by smartphones, and as we look ahead to 2020, predictions indicate a 53 percent compound annual growth rate (CAGR) in mobile data traffic, attaining a total 30 exabytes/mo globally.

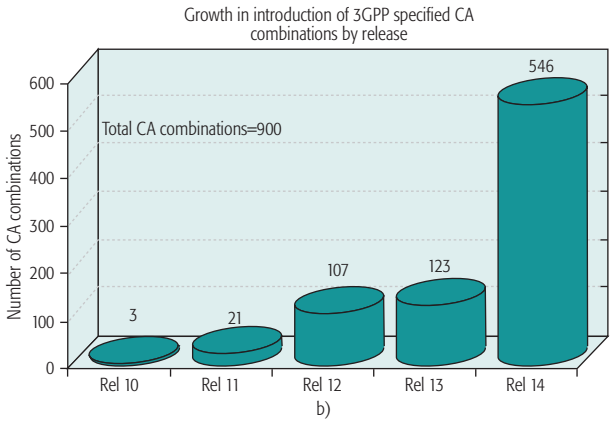
LTE-ADVANCED PRO: SOLUTIONS FOR THE CHALLENGE OF BIG MOBILE DATA

In order to address this explosive demand for data rates and total mobile data consumption, manufacturers are called to increase data throughput of consumer UE. A number of enabling features are being standardized and rolled out in commercial handset products. The highest priority to date has been deployment of carrier aggregation (CA), which was introduced in the Third Generation Partnership Project's (3GPP's) Release 10, and involves the addition of more and more carrier bandwidth. CA essentially allows mobile operators to "widen the pipe" and enable higher data rates simply by the simultaneous use of more spectrum as a dedicated resource to a single user. LTE is defined to support flexible channel bandwidths from 1.4 MHz to a maximum of 20 MHz, but these critical extra channels (each up to 20 MHz wide) can be added within a defined band of operation (intra-band CA) or in additional different bands of operation (inter-band CA). The number of combinations of the channel allocations and combinations of bands employed for CA in the standard has exponentially grown over the last several years, as indicated in the summary by the 3GPP Release in Fig. 1, and we see the continued use of CA as a vital part of increasing data rates for consumers [2]. This feature is further illustrated in Fig. 2, where the addition of component carriers (CCs) that aggregate more bandwidth to the signal can benefit users throughout the entire cell (all the way to cell edge). The darker shade of the larger number of aggregated

Digital Object Identifier:
10.1109/MCOM.2017.1601221

The authors are with Skyworks Solutions, Inc.





Band groups	2DL CA combinations
LB/MB	5/2, 5/66, 12/2, 12/66, 13/2, 13/66, 29/2, 29/66
LB/HB	5/7, 12/7, 5/30, 12/30, 29/30
LB/LAA	5/46, 12/46, 13/46
MB/MB	2/4, 2/66, 2/2, 4/4, 25/25, 66/66
MB/HB	2/30, 2/7,m 4/30, 66/7
MB/LAA	2/46, 66/46
HB/LAA	7/46, 41/46

Band groups	3DL CA combinations
LB/LB/LB	5/12/12
LB/LB/MB	2/12/12, 2/5/5, 4/12/12, 4/5/5, 5/5/66
LB/LB/HB	5/5/30
LB/MB/MB	12/66/66, 13/66/66, 2/12/66, 2/13/66, 2/5/66
LB/HB/HB	5/7/7
LB/LAA/LAA	13/46/46, 5/46/46
MB/MB/MB	2/66/66
MB/MB/HB	2/7/66
MB/HB/HB	2/7/7, 4/7/7
MB/LAA/LAA	2/46/46, 4/46/46, 46/46/66
HB/LAA/LAA	7/46/46

Band groups	4DL CA combinations
LB/LB/LB/MB	2/5/12/12, 4/5/12/12
LB/LB/MB/MB	2/2/12/12, 2/2/5/5, 2/4/12/12, 2/4/5/5, 2/5/5/66, 4/4/12/12,
LB/LB/MB/HB	2/5/5/30, 4/5/5/30
LB/MB/MB/MB	13/66/66/66, 2/13/66/66, 2/2/13/66, 2/2/5/66, 2/5/66/66,
LB/HB/LAA/LAA	5/7/46/46
LB/LAA/LAA/LAA	13/46/46/46, 28/46/46/46, 5/46/46/46
MB/MB/MB/MB	2/2/66/66, 2/66/66/66
MB/MB/HB/HB	2/4/7/7
MB/LAA/LAA/LAA	2/46/46/46, 4/46/46/46, 66/46/46/66
HB/LAA/LAA/LAA	7/46/46/46

Band groups	5DL CA combinations
LB/LB/MB/MB/HB	2/2/5/5/30, 2/4/5/5/30, 4/4/5/5/30
LB/HB/LAA/LAA/LAA	5/7/46/46/46
LB/LAA/LAA/LAA/LAA	5/46/46/46/46
MB/LAA/LAA/LAA/LAA	2/46/46/46/46, 4/46/46/46/46, 66/46/46/46/46
HB/LAA/LAA/LAA/LAA	7/46/46/46/46

LB: B12, B13, B29, B5, B26
MB: B2, B25, B4, B66
HB: B7, B30, B41
License assisted access (LAA): B46

c)

Figure 1. a) Mobile data traffic per month and traffic growth predictions 2015–2020 (1 exabyte = 1018 bytes) [1]; b) exponential growth in the definition of band combinations employed for carrier aggregation as part of the 3GPP standard [2]; c) North America example of requirements for downlink CA combinations across 2DL, 3DL, 4DL, and 5DL use cases.

CCs indicates higher throughput as this feature linearly increases data rate proportional to the total bandwidth employed.

Another technique designed to increase the spectral efficiency of bandwidth is to effectively increase the data rate in bits per Hertz. Termed "higher order modulation," defined in 3GPP's Release 12 (spring 2015) to be a maximum of 256-quadratuer amplitude modulation (QAM) for the downlink (DL), and 3GPP's Release 14 (expected spring 2017) to support a maximum of 256-QAM for the uplink (UL). As the standard has started with modulations of quadrature phase shift keying (QPSK) (2 bits/symbol), to 16-QAM (4 bits/symbol), to 64-QAM (6 bits/symbol), and now to 256-QAM (8 bits/symbol), the spectral efficiency is increased by the factor of bits per symbol. This increase in bits/symbol requires a correspondingly higher signal strength or signal-to-noise ratio (SNR), and closer proximity to the eNodeB as shown in Fig. 2.

The other new technique being implemented extends the number of data streams to increase data rates. The application of multiple-input multiple-output (MIMO) spatial multiplexing effectively transmits multiple data streams (or layers) from a number of antennas at the transmitter to multiple antennas on the receiver. This application uses the spatial differences of the antenna reception and multi-path through varying radio environments of each data stream in order to separate out the overlying signals even though they are transmitted at the same frequency. This digital extraction of the signals based on known unique radio path transfer functions (derived from reference signals within each link) enables a further multiplication factor of the data rate according to the number of transmit/receive antennas that are employed. As an example of the DL signals, if four data streams are transmitted from the base station (eNodeB) and four separated antennas with low envelope correlation coefficient are used for reception at the UE handset, this 4 × 4 DL MIMO link will be able to support two times the data rate of a 2 × 2 DL MIMO link (two antennas at the eNodeB and two antennas at the UE) and four times a single (1 × 1, or single-input single-output [SISO]) antenna reception. The application of MIMO requires SNR to function adequately and may

The technology can also be used to boost the SNR by transmitting additional copies of the same data stream and using the multiple receiving antennas to decompose the same effective data stream using pre-coding and the difference in radio environments of each antenna to improve the reception of that one data stream.

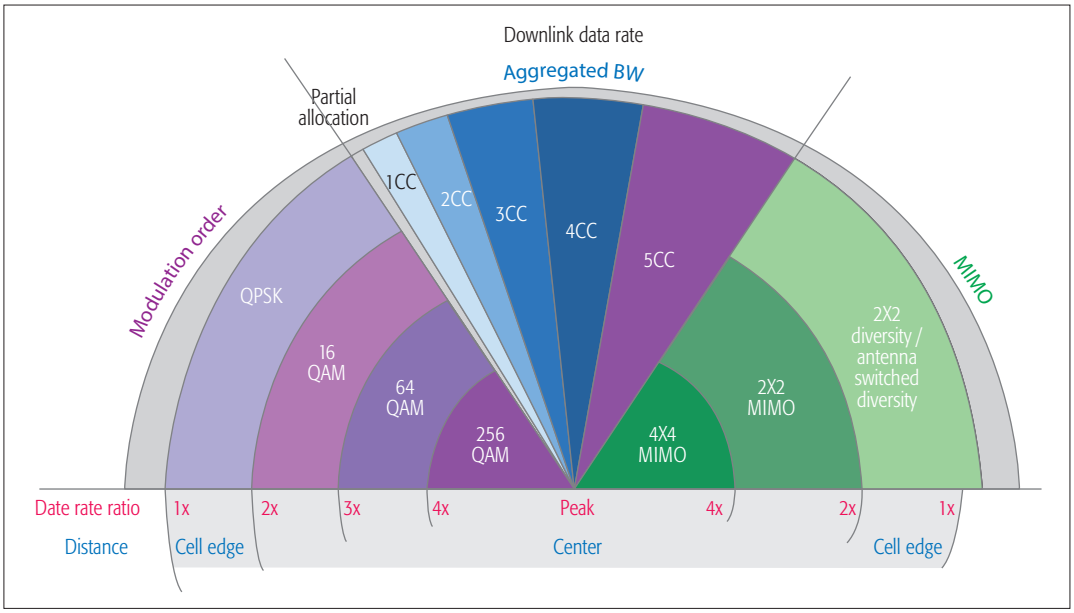


Figure 2. Downlink LTE-Advanced features and impact on data rate throughout the cell.

require stronger signals with less interference and closer proximity to the eNodeB than a corresponding lower data rate SISO operation, as also demonstrated in Fig. 2.

The technology can also be used to boost the SNR by transmitting additional copies of the same data stream and using the multiple receiving antennas to decompose the same effective data stream using pre-coding and the difference in radio environments of each antenna to improve the reception of that one data stream. The SNR of a single (SISO) reception can effectively be doubled when the same data stream is encoded, transmitted from two separate antennas at the eNodeB, and received by two antennas on the UE, providing 3 dB in "diversity gain." This concept of diversity can be further applied to four antennas for yet another doubling of SNR (or 3 dB more increase) for the same data stream, and an extended range or extended distance from the eNodeB. In a 2015 study of the impact of 4 × 4 MIMO on DL data rates and coverage in their B41 network, U.S.-based carrier Sprint demonstrated large increases in throughput across a large SNR range, effectively improving data rates 50–60 percent at cell edge through SNR gains and diversity gain, and leveraging sufficient SNR to improve throughput 45 percent at mid-cell and 22–38 percent at cell center [3]. Similar studies by Orange in 2012 indicated a 60 percent increase in average throughput in upgrading from 2 × 2 to 4 × 4 DL MIMO [4], and more recently SK Telecom indicated a 42 percent average throughput increase with 4 × 4 MIMO vs. 2 × 2 [5]. This technology of packing more bits into the existing spectrum is extremely attractive to the operators, who are required to pay so much for the fundamentally limited resource of that available spectrum. The increase of throughput throughout the cell and improvement all the way to the critical cell edge user experience at the outer extent of Fig. 2 is part of the reason behind the rapid adoption of 4 × 4 MIMO on the DL of higher tier handsets.

This diversity transmission mode and the use of multiple antennas for spectral efficiency, robustness against multipath, diversity benefit, and SNR improvement is such a powerful concept that the LTE standard mandated that all UEs must have at least two active simultaneous receive antennas for any given operation in any given band, and as we add bands in CA, each is similarly added using two simultaneously active receivers. A further modification of this concept of diversity also includes "antenna switched diversity," effectively selecting from a number of available antennas to choose the best of those, and operating from fewer antennas with less current consumption but with better overall performance because of that enabling choice.

In describing the coverage of these DL features throughout the cell, as shown in Fig. 2, it is actually the case that cell edge performance in LTE-Advanced (LTE-A) is most often limited by *UL* power from the UE [3]. Support for even higher data rates in the UL force the spreading of limited UL power across a wider spectrum in UL CA, causing further decreases in the individual power per resource block in dBm per Hertz. This is compounded by the challenge of meeting emissions requirements and necessary transmission at lower total powers from the UE as the transmit spectrum widens on the UL. However, higher data rates inherently require elevated SNR and signal quality. As seen in Fig. 2, closer to the base station, uplink power to preserve the link becomes less of a limitation, and the priority for higher DL data rate and the DL SNR become the limiting factors. With carrier networks' attention on DL video driving enhanced mobile broadband (eMBB), there is an increasing focus on enhancements to the DL.

To attain faster data rates and improve network efficiency, mobile network operators and device manufacturers employ multiple combinations of these advanced features in LTE-Advanced Pro. For example, the calculation of the resulting data rate for a standard LTE DL signal employ-

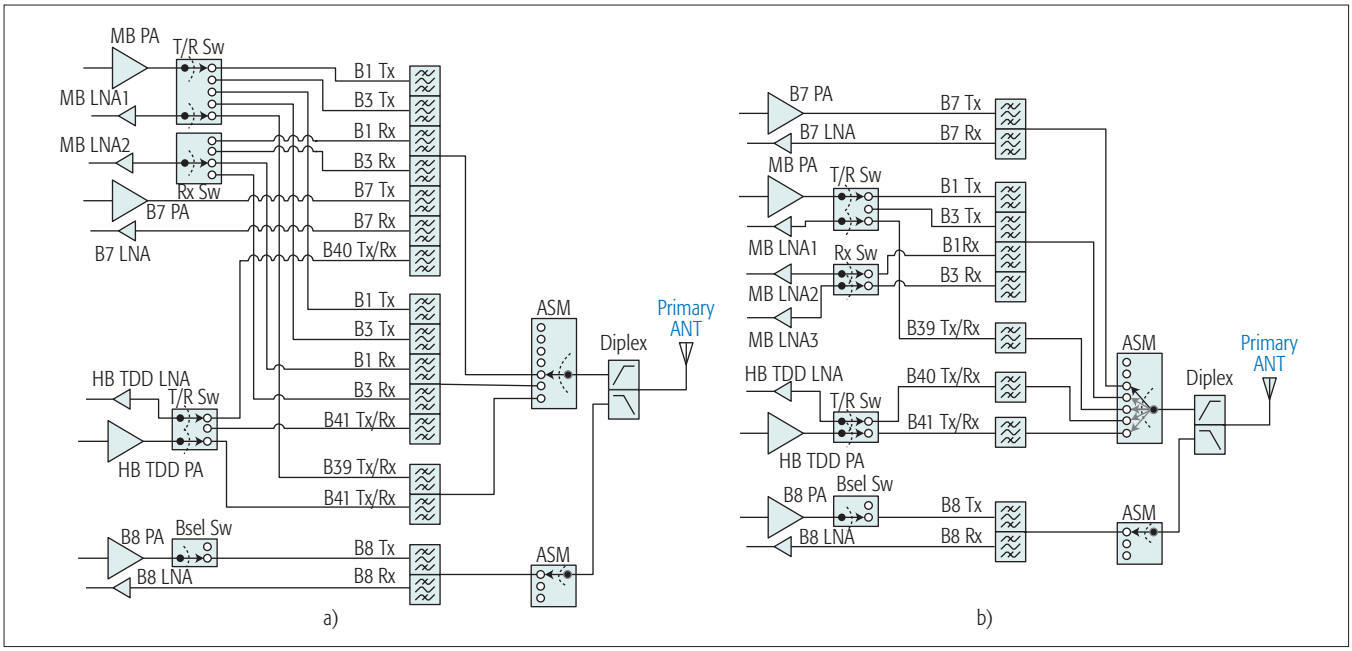


Figure 3. Example implementation options for primary PA+Duplexer+ASM module (PAiD) carrier aggregation support of B8/B1, B8/B3, B8/B7, B1/B3/B7, B1/B3/B40, B7/B40, B1/B41, B3/B41, and B39/B41: a) permanently ganged *N*-plexer filter configuration; b) more optimal switch-combined approach having lower loading loss and less filter duplication.

ing diversity gain for a single data stream at cell edge in the outer extent is shown in Fig. 2. For this demonstration, we consider a single QPSK 20 MHz DL channel, where each subframe is 1 ms long and consists of two time slots with seven symbols each. Each of those QPSK symbols consists of 2 bits, with 20 MHz bandwidth containing 1200 such 15 kHz resource element subcarriers for a total of 33.6 Mb/s. About 75 percent of these bits are useful data, while around 25 percent are overhead due to the physical DL control channel (PDCCH), reference signal, synchronization signals, PBCH, and some coding. This brings our QPSK 20 MHz DL SISO baseline peak data rate at the outer extent of the cell to 25.2Mb/s. When we apply the enhanced features previously described nearer the eNodeB with sufficient SNR and signal quality available, the benefit of 256-QAM (factor of 4 increase), 4 × 4 DL MIMO (factor of 4 increase), and 5 × 20 MHz channel CA (factor of 5 increase), all multiply to a resulting peak data rate of 2.016 Gb/s. LTE-Advanced Pro enables greater than 1 Gb/s data rates in the UE, and the complexities of network density, feature support, and UE performance determine the coverage area and distance from the eNodeB, as well as robustness of the achievable data rates throughout the cell [6].

Specific Architecture and Performance Challenges of Downlink Carrier Aggregation

As the explosion in number of required CA combinations in Fig. 1 shows, the RFFE must support a large number of complex simultaneous RF paths. When considering DL CA where a single transmit signal is combined with one or more paired receive channels, two of the more challenging use cases are those that fall into the following categories:

- A harmonic of the transmit channel falls directly onto one of the active receive channels.
- The inter-band CA involves two signal bands whose Tx and/or Rx passbands are relatively close together.

For the first case, where the harmonic of a lower frequency band falls into a CA partner receive band, there is an example shown in Fig. 3 depicting the specific cases of B8/B3 (2nd harmonic of B8 Tx 880–915 MHz falls into the B3 Rx band 1805–1880 MHz) and B8/B7 (3rd harmonic of B8 Tx 880-915 MHz falls into the B7 Rx band 2620–2690 MHz). The harmonic levels (around –10 dBm) are significantly higher than the typical noise of the transmitter and must be attenuated to a level below –85 dBm before the low band input of the diplexer in order to avoid desensing the B3 and/or B7 primary/diversity receivers. Multiple additional isolations within the front-end must be well below this challenging attenuation of the conductive path. Overall harmonic management is a difficult balance of shielding, distributed low-loss harmonic filtering, and grounding for optimal isolation that is critically aided by integration and proper partitioning of PA+Duplexer+Switch module packages (PAiDs). The second primary challenge is related to the merging of closely spaced bands, and an example of this is also shown in the two example implementations in Fig. 3. On the left in Fig. 3a, closely spaced bands are permanently ganged together in large groups, so-called *N*-plexer filter arrays, demonstrated here with a 7-plex to deliver B1/B3/B7, B1/B3/B40, and B7/B40, a 5-plexer to deliver B1/B41 and B3/B41, and a diplexer to deliver B39/B41. This approach is a common brute force architecture that leverages a fixed set of specific CA combinations, and enables less calibration for the fewer possible RF path configurations, but without co-design with the antenna switch to enable

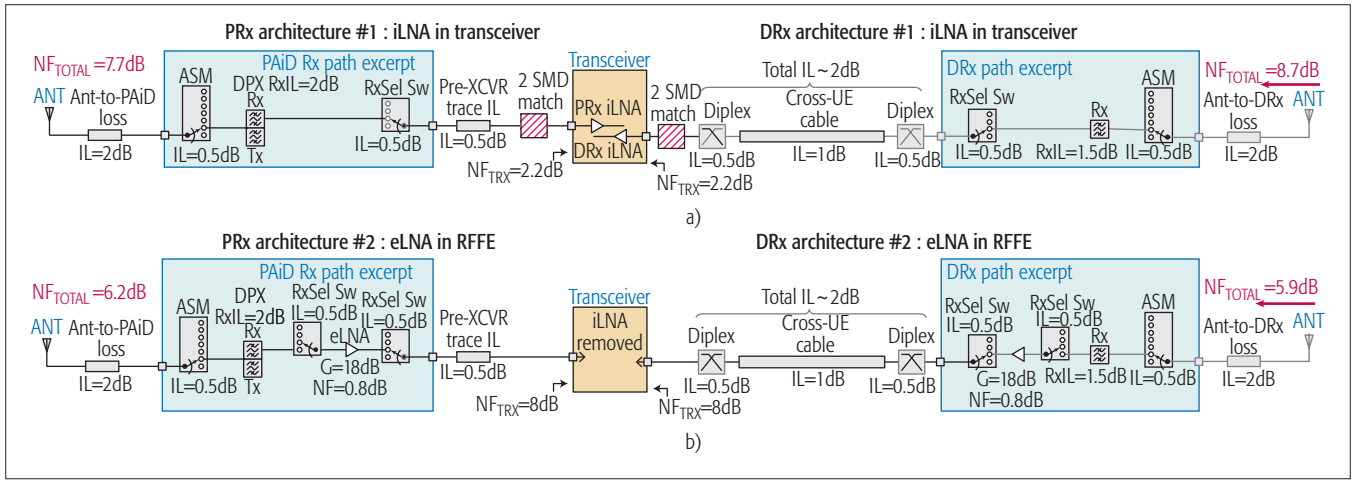


Figure 4. Receiver link budget for: a) RFIC integrated LNA; b) external LNA in the RFFE, demonstrating 1.5 dB lower NF for the eLNA solution in the primary receiver, and 2.8 dB lower NF for the eLNA solution in the diversity receiver.

reconfiguring or switching filter combinations in and out using the switch. The increasing loading losses as more filters are ganged, along with the inflexible configuration to address additional CA, is compounded here by the inability to gang filters whose passbands overlap, such as B39/B1/B3 and B7/B41. In order to deliver all of these combinations, filters need to be duplicated in the ganged arrangements at some cost and area penalty. In contrast, the solution on the right employs a flexibly configured switch able to simultaneously engage two arms to connect and join various combinations of filters for different CA pairings (e.g., the ASM switch arms in Fig. 4b connecting to both the B1B3 quadplexer and the B7 duplexer to achieve B1/B3/B7 3DL CA) . This not only enables all of the specific CA combinations to be delivered, but also eliminates five filters compared to the ganged approach. This also exhibits much lower insertion loss when single band operation is engaged because no additional filter loading is switched in. This advantage of better single band operation is a critical factor for the cell edge user experience, perhaps the most common and most important priority of field use cases. When the filters are switched in and simultaneously conjoined to a common RF path, their loading and relative impedances both in-band and out-of-band must be managed extremely carefully in an integrated design. This is similar to the permanent ganging example in Fig. 3a, but the difference is that the loading loss is only suffered when in CA operation. As the number of CA combinations involving overlapping bands continues to increase according to Fig. 1, flexibly switch-combined architectures will be preferred for performance and cost consideration. The primary receiver block diagrams shown in Fig. 3 of course need to be supplemented with the mandatory additional receive chains to support receiver diversity, along with the additional receivers required for higher order MIMO on the DL in the complete phone solution. Figure 3 depicts only the primary (one of the four) to focus in on the significant challenges of supporting both Tx and Rx on that primary antenna feed, but for 4 × 4 DL MIMO support, there must be 4 active receivers on 4 dedicated antennas, as described later in Fig. 5.

Optimization of RF Front-End Receiver Architectures

The historical partitioning and implementation of the transceiver RF integrated circuit (RFIC) and the RFFE is shown in Fig. 4a. Transceiver design and interface with the front-end is complicated by demand to support the exploding number of bands and CA combinations, along with the sheer number of simultaneous transmit and receive chains required. Originally, differential receiver inputs were employed to make full use of common-mode rejection and leverage the advantages of limited voltage swing and limited headroom against aggressively shrinking complementary metal oxide semiconductor (CMOS) gate dimensionality and associated lower supply and breakdown voltages. As the number of transmitter and receiver pins started to grow, the shrinking CMOS supply voltages started to limit the actual common mode rejection benefits due to requirements for pseudo-differential implementations (which are not fully differential with shared tail currents). At the same time, the die/transceiver solution size started to become fundamentally limited by the number of pins, and the required matching networks for differential interfacing became too costly in PCB phone board layout space. It became clear that these receiver RF interfaces needed to migrate to become single-ended, and so they have. Differential receive interfaces on the frequency-division duplex (FDD) filters gave way to single-ended interfaces, and acoustic filter manufacturers found ways to continue to improve the smaller filter's isolation and insertion loss despite the change.

LTE's introduction with Release 8 of the 3GPP standard in 2008 required that receiver diversity be employed as a mandatory requirement (2 × 2 DL with two antennas at the eNodeB and two antennas at the UE receiver), doubling the number of active receivers. Transmission modes were defined to leverage the capability of full 2 × 2 DL MIMO for data rate advantage in high SNR radio environments, as well as the diversity gain benefits of 2 × 2 Rx diversity gain at cell edge to overcome fading multipath and extend the range of the DL signal connection. LTE-A's intro-

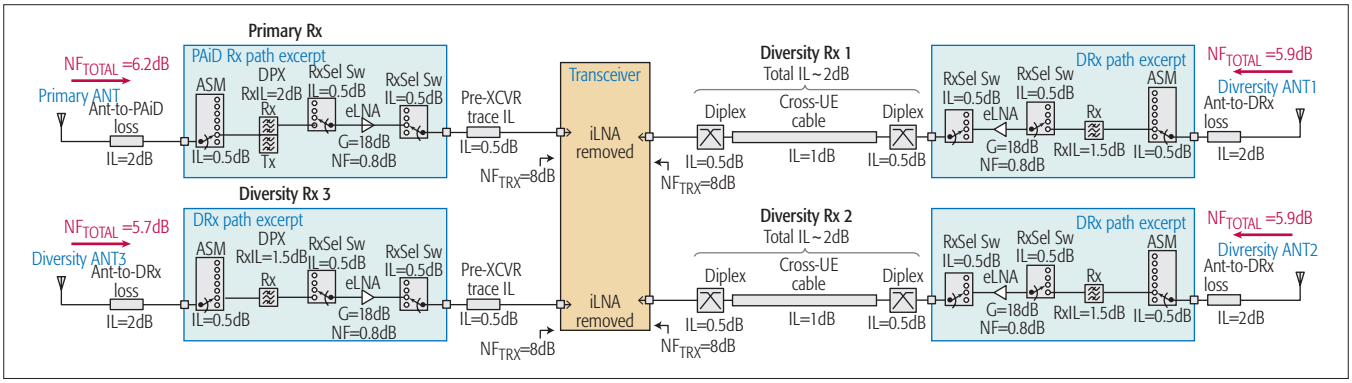


Figure 5. Receiver antenna connectivity and link budget for 4 × 4 MIMO DL support.

duction in Release 10 of DL CA-enabled summation of simultaneous component DL carriers through simultaneous and separate receive paths significantly increased the available spectrum and data throughput for each individual user. This constrains the number of physical receiver paths needing to be increased to support concurrent use. This effectively means that paths could not be reused because they were both active at the same time carrying different signals that needed to be conditioned independently of one another. LTE-Advanced Pro, added in Release 13, is the natural extension of this concept, in the form of an optional feature to support 4 × 4 DL MIMO, which again doubles the number of potentially simultaneous active receive chains.

In order to maintain the required orthogonality and low envelope correlation coefficient in the handset, the physical location of these separate antennas requires relatively large trace losses and cross-UE cable insertion losses to get back to the transceiver where the LNA inputs were located. It became clear that in order to optimize performance, the switching, filtering, and LNA need to be as close as possible to the physical antenna. Once the signal is amplified with low noise figure (NF) in the LNA, post-LNA loss in both the signal and elevated noise level have less challenge against the thermal noise floor, and the overall SNR is preserved despite the extra post-LNA losses. When the diversity antenna is on the opposite side of the UE, as shown in Fig. 4a, the cross-UE trace and/or cable losses can be in excess of 1–2 dB, and this adds directly to the overall NF as a direct penalty to Rx sensitivity. If the LNA is placed remotely, close to the antenna as shown in Fig. 4b, the loss before the LNA is minimized. The noise figure impact due to loss after the LNA is reduced by the amount of that gain, typically a linear reduction factor of ~ 40–65. As illustrated in the example of Figs. 4a and 4b, the NF reduction between the architecture with the LNA in the transceiver and loss between transceiver and the antenna vs. the LNA placed remotely close to the antenna with less insertion loss before the LNA is 2.8 dB for this diversity receiver case.

It is primarily for this performance benefit that Rx diversity modules have been developed to be placed as close as possible to the antenna. Some additional benefit is gained from the facts that the external LNA matched specifically to the integrated Rx filter can show much lower noise figure (roughly 0.8 dB vs. 2 dB at 2 GHz), and all of the surface mount components required for input matching of the LNA are integrated in the module, no longer taking up space on the phone PCB.

For the primary receiver, a similar analysis shows incremental benefits as a function of the architecture and LNA improvements such that 1.5 dB improvement in Rx sensitivity can be achieved with an external LNA (eLNA) in the RFFE, as opposed to an LNA integrated in the transceiver, as demonstrated in the left portion of Fig. 4. This performance advantage alone is compelling, but is supplemented by the benefits of not requiring any matching components between the Rx path in the RFFE and the transceiver input, reducing the cost and area required on the phone board. The primary receiver also faces more challenges in the rejection of the Tx carrier power leakage onto the active primary receive path than does the diversity receiver, which benefits from the antenna isolation. Differences like these between primary and diversity receive drive slightly different filter attenuation requirements and the associated extra insertion loss that comes with higher out-of-band attenuation, and are a large part of optimizing the components as configured in Fig. 4b.

When considering the connectivity of the front-end to support 4 × 4 MIMO DL, four separate antennas with low envelope correlation are required, and typically are designed for maximal isolation and physical separation in the four corners of the UE. The requirement for four good antennas with similar radiated performance is a significant challenge given the volume constraints for reasonable radiation efficiency and the typically thin metal chassis form factor of modern smartphones. Support for the lowest frequency bands is the most difficult, where antenna aperture tuning and priority are employed to salvage the extremely narrowband radiation efficiency of the lower frequency radiators/exciters. No more than two antennas supporting lower frequencies below 960 MHz are possible, and thus only bands above 1.7 GHz are considered viable for 4 × 4 MIMO feature support in modern form factor UEs. An interesting aspect of the antenna configuration is the requirements for two antennas supporting lower frequency, shared for > 1.7 GHz cellular support as well. With two additional antennas supporting > 1.7 GHz, all as orthogonal as possible with low envelope correlation coefficient, this configuration drives a common antenna system of the four antennas that tends toward a common antenna interface of four feeds, as depicted in Fig. 5.

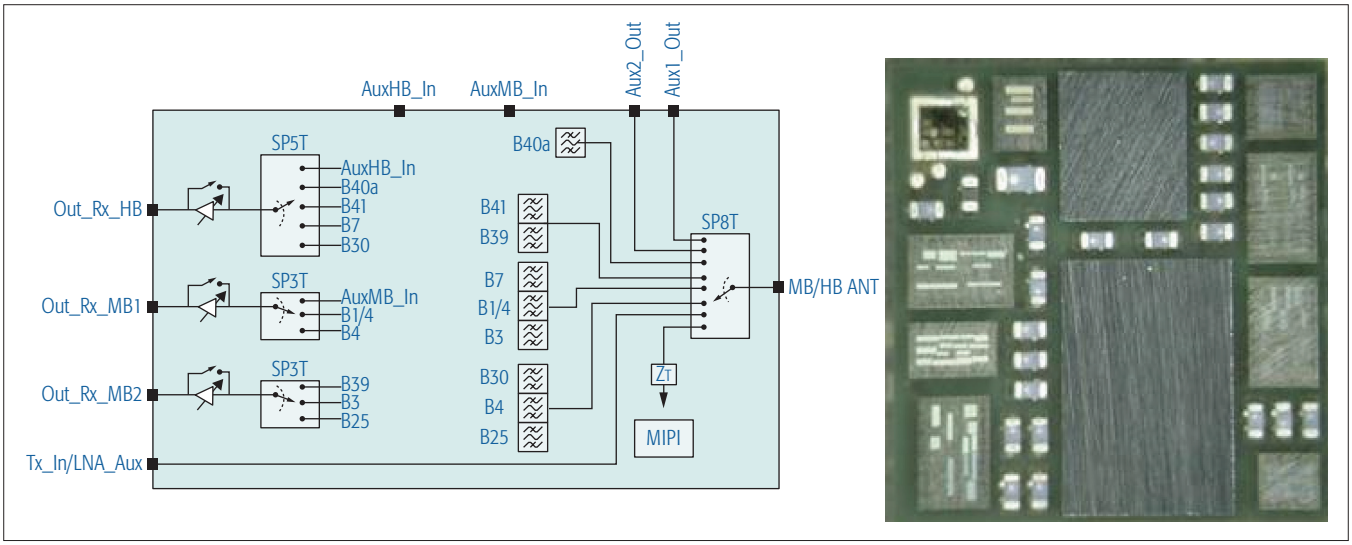


Figure 6. Diversity receiver module SKY13750 supporting B1/B25/B3/B4/B39 (mid bands) and B30/B40/B7/B41 (high bands), and a module photograph.

Whereas previous implementations across original equipment manufacturers (OEMs) in support of LTE had a range of two- to four-antenna solutions, going forward support of 4 × 4 DL MIMO forces a more converged four-antenna solution across most smartphones. The remote placement of the LNA and corresponding integrated modules as close as possible to the four antenna feeds are critical to reduce loss and overall noise figure, and the RFFE is depicted in Fig. 5.

Diversity Receive Integrated Modules for CA and Higher Order MIMO

The design of these advanced diversity receive modules requires multiple technologies optimized for switching, acoustic filtering, and active LNAs, which must be co-designed to leverage the benefits of hybrid assembly in multi-chip module packaged integration. The filter itself is specifically matched to the input impedance of the LNA, minimizing trace loss and other matching transformation insertion losses for the lowest noise figure. Thus, managing out-of-band attenuation requirements, all with careful co-design of other filters that may be switch-combined in CA pairing within the same module as described earlier, is important. The discrete solution is unable to switch-combine filters in flexibly programmed CA pairings due to long trace losses and phase shifts on the phone PCB, and the overall discrete solution is commonly twice as large as the integrated module containing comparable band support. As more bands become required, the size advantage of the integrated solution will become even greater. The higher frequency bands (> 1.7 GHz) within the UE are all candidates to support 4 × 4 MIMO on the DL. However regional operator demand for the feature and whether the UE is designed as a global smartphone to support all regional requirements will determine the number of bands, and which ones, are populated to support 4 × 4 MIMO. An example of a global diversity receive module is shown in the block diagram and module photograph of Fig. 6 that supports B1/B25/B3/B4/B39 (mid bands) and B30/B40A/B7/B41 (high bands) and all associated globally required CA combinations. This module serves as a CA-capable MB/HB diversity receive module, but can also be placed additionally to support 4 × 4 MIMO in the DL of these same bands with connection to the other available antenna feeds, as shown in the RFFE architecture of Fig. 5.

Conclusion

The incredible demand for mobile data capacity, ever rising data rates, and higher quality user experience throughout the cell is driving very complex features into modern smartphones. LTE-Advanced Pro is answering the call, enabling expanded bandwidth in the form of CA, increased spectral efficiency of that bandwidth by employing higher order modulations, and higher order MIMO techniques. Critical aspects of spectral efficiency to make the best possible use of the limited spectrum resource and significantly improve throughput throughout the entire cell are compelling reasons for the accelerating demand for 4 × 4 MIMO. The emphasis on DL presented here is simply to address the predominant asymmetry of present networks for download of video and other content. Fundamental limitations of the networks based on uplink power have also been described. In order to keep up with the exponential growth in mobile data, concurrent application of these features of CA, higher order modulation, and 4 × 4 MIMO must be used. Supporting each of these features is a challenge for the RF front-end, but complications of insertion loss, noise figure, isolation, and self-desense are further compounded when they are all engaged simultaneously. Architectures on the primary path to address DL CA challenges, trade-offs of ganged filtering vs. switched combined filter topologies, advantages of LNA placements closer to the antenna, architectures to support 4 × 4 MIMO, and a specific example of a MIMO and CA-capable diversity receive module have been described. Both transmit and receive, across all antenna connectivity, and incorporating all the capabilities and limitations of the transceiver and modem must be considered in a holistic system perspective to

address these complex RF subsystem challenges for next generation handset implementations.

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020," 3 Feb. 2016; http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html; accessed Dec. 1, 2016.

[2] 3GPP 36.101 "Evolved Universal Terrestrial Radio Access (E-UTRA) User Equipment (UE) Radio Transmission and Reception Specification, Rel. 14.2.1, Jan. 14,2017.

[3] H. Sava, "LTE-Advanced, Higher Order MIMO, CA, and Increased UL Tx Power," *Proc. IWPC Wksp.*, Madrid, Spain, May 11–13, 2015.

[4] J.-B. Landre, Z. El Rawas, and R. Visoz, "Realistic Performance of LTE: In a Macro-Cell Environment," *Proc. IEEE VTC-Spring*, 2012 Yokohama, Japan, 2012, pp. 1–5.

[5] Y. Kim *et al.*, "Performance Analysis of LTE Multi-Antenna Technology in Live Network," *Proc. URSI Asia-Pacific Radio Science Conf.*, Seoul, Korea, 2016, pp. 1302–05.

[6] GSMA, "Unlocking Commercial Opportunities: From 4G Evolution to 5G," Feb. 1, 2016; http:/www.gsma.com/network2020; accessed Dec. 1, 2016.

Biographies

David R. Pehlke [SM] (David.Pehlke@skyworksinc.com) is currently a senior technical director of Systems Engineering at Skyworks Solutions. He received his Ph.D. and M.S.E. in the areas of solid-state device physics and technology optimization of III-V compound semiconductors from the University of Michigan and his S.B.E.E from MIT. Previous work experience includes the Rockwell Science Center, Ericsson Mobile Platforms, Silicon Laboratories and ST-Ericsson, and Skyworks. He presently chairs the IEEE Buenaventura Communications Society Chapter.

Kevin Walsh is currently a senior director of Mobile Product Marketing at Skyworks Solutions. He received a B.S.E.E. in microwave engineering and solid state semiconductors from the University of Massachusetts with advanced technical marketing work with Worcester Polytechnic Institute and Caltech. He has gathered extensive marketing experience in mobile systems from experience at IBM Semiconductor, Siemens Microelectronics, RF Micro Devices, and Skyworks Solutions. He has responsibility for long-term product roadmap and mobile operator engagements, and is working on moving products into the emerging 5G ecosystem.

Integrated Full Duplex Radios

Jin Zhou, Negar Reiskarimian, Jelena Diakonikolas, Tolga Dinc, Tingjun Chen, Gil Zussman, and Harish Krishnaswamy

Full duplex wireless has drawn significant interest in the recent past due to the potential for doubling network capacity in the physical layer and offering numerous other benefits at higher layers. However, the implementation of integrated full duplex radios is fraught with several fundamental challenges.

ABSTRACT

Full duplex wireless has drawn significant interest in the recent past due to the potential for doubling network capacity in the physical layer and offering numerous other benefits at higher layers. However, the implementation of integrated full duplex radios is fraught with several fundamental challenges. Achieving the levels of self-interference cancellation required over the wide bandwidths mandated by emerging wireless standards is challenging in an integrated circuit implementation. The dynamic range limitations of integrated electronics restrict the transmitter power levels and receiver noise floor levels that can be supported in integrated full duplex radios. Advances in compact antenna interfaces for full duplex are also required. Finally, networks employing full duplex nodes will require a complete rethinking of the medium access control layer as well as cross-layer interaction and co-design. This article describes recent research results that address these challenges. Several generations of full duplex transceiver ICs are described that feature novel RF self-interference cancellation circuits, antenna cancellation techniques, and a non-magnetic CMOS circulator. Resource allocation algorithms and rate gain/improvement characterizations are also discussed for full duplex configurations involving IC-based nodes.

INTRODUCTION

One of the long-held precepts of wireless communication has been that it is impossible for a wireless device to transmit and receive at the same time and at the same frequency because of the resulting self-interference (SI). Recent efforts have challenged this wisdom, opening the door to *full duplex* (FD) wireless, which has the potential to immediately double network capacity at the physical (PHY) layer and offers many other benefits at the higher layers.

The concept of FD communication is certainly not entirely new. The earliest pre-electronic telephone handsets used hybrid transformers to isolate the earpiece from the microphone, enabling FD communication on a two-wire loop to the central office. In the realm of wireless, the Plessey Groundsat system of the 1970s was a military-specification FD wireless system capable of operating over radio channels within the 30–76 MHz VHF band.

However, achieving FD operation in commercial wireless applications, such as cellular communications and WiFi, is fraught with several challenges. The fundamental challenge associated with FD wireless is the tremendous transmitter (TX) SI, or echo, that appears at the receiver (RX). This SI can be anywhere between 90–120 dB (*a billion to a trillion times*) more powerful than the desired signal depending on the application. This powerful SI is further susceptible to the uncertainties of the wireless channel (e.g., frequency selectivity and time variance) and the imperfections of the transceiver electronics (nonlinear distortion and phase noise, to name a couple). These challenges are further exacerbated when integrated implementations targeting cost-sensitive and form-factor-constrained mobile devices are considered. Finally, to fully utilize the benefits of FD communication, wireless systems will require a fundamental rethinking of not only the PHY layer but also the medium access control (MAC) layer, and a careful co-design of the two.

Initial research performed a few years ago established the feasibility of SI cancellation and FD operation in commercial wireless applications using laboratory bench-top equipment and off-the-shelf components [1]. More than 100 dB SI cancellation has also been demonstrated in [2] for military applications. However, the self-interference cancellation (SIC) techniques utilized in these works are fundamentally not compatible with small-form-factor/integrated circuit (IC) implementations. More recently, research on integrated FD radios and associated SI cancellation techniques has emerged [3–7].

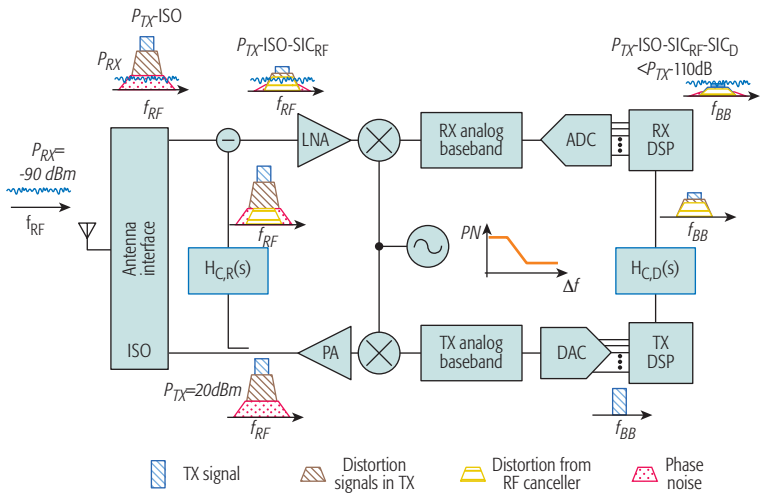
This article reviews recent research at Columbia University on integrated FD radios spanning RF, analog and digital SIC, FD antenna interfaces, and non-magnetic complementary metal oxide semiconductor (CMOS) circulators that enable single-antenna FD [3–5, 8, 9]. This article also touches on the FlexICoN project at Columbia, which is taking a holistic cross-layered approach to develop FD radios and networks from PHY to MAC. It covers resource allocation algorithms and rate gate/improvement characterizations for FD configurations involving IC-based FD nodes.

CHALLENGES ASSOCIATED WITH INTEGRATED FULL DUPLEX RADIOS

Figure 1 depicts the block diagram of an FD radio that incorporates antenna, RF, and digital SI suppression. The indicated transmitter and minimum

Jin Zhou is with the University of Illinois at Urbana-Champaign; Jelena Diakonikolas is with Boston University; Harish Krishnaswamy, Gil Zussman, Negar Reiskarimian, Tolga Dinc, and Tingjun Chen are with Columbia University.
This research was performed while Jin Zhou and Jelena Diakonikolas were at Columbia University.

Digital Object Identifier:
10.1109/MCOM.2017.1600583



		Specifications	
System definition	TX average output power (P_{TX})	20 dBm	
	Signal bandwidth	20 MHz	
	RX noise floor (P_{RX})	-90 dBm	
	Total SIC	110 dB	
SIC in the antenna and RF domains	Antenna isolation (ISO)	30 dB	50 dB
	RF SIC (SIC_{RF})	20 dB	20 dB
Analog/RF front-end circuit performance	DR_{ADC}*	72 dB	52 dB
	RX effective IIP_3#	-3 dBm	-23 dBm
	RF canceller OIP_3#	-3 dBm	-23 dBm
	VCO phase noise @ 1MHz offset$	-106 dBc/Hz	-86 dBc/Hz
Digital SIC budget	Digital SIC - linear	60 dB	40 dB
	Digital SIC - IM3	60 dB	40 dB

*12 dB margin has been added to DR_{ADC}.
#Assume the SI-induced IM3 at the ADC input has the same power level as the residual SI after RF SIC.
$Assume 10ns delay in the SI channel.

Figure 1. Block diagram of an FD radio featuring antenna, RF and digital SI suppression, along with a depiction of the various transceiver non-idealities that must also be managed for effective FD operation.

received signal power levels are typical of WiFi applications. Also depicted are various transceiver non-idealities that further complicate the SI suppression problem.

Achieving > 100 dB SI Suppression

The power levels indicated in Fig. 1 necessitate > 110 dB SI suppression for WiFi-like applications. Such an extreme amount of cancellation must necessarily be achieved across multiple domains (here, antenna, RF and digital), as > 100 dB precision from a single stage or circuit is prohibitively complex and power inefficient. The suppression must be judiciously distributed across the domains, as suppression in one domain relaxes the dynamic range requirements of the domains downstream. Furthermore, all cancellation circuits must be adaptively configured together — optimization of the performance of a single cancellation stage alone can result in residual SI that is sub-optimal for the cancellers downstream.

Transceiver Non-Idealities

The extremely powerful nature of the SI exacerbates the impact of non-idealities such as nonlinearity and phase noise, particularly for IC implementations. For instance, nonlinearity along the transmitter chain will introduce distortion products. Antenna and RF cancellation that tap from the output of the transmitter will suppress these distortion products, but linear digital cancellation will not as it operates on the undistorted digital signal. Depending on the amount of antenna and RF cancellation achieved, the analog receiver front-end may introduce distortion products as well, as may the RF cancellation circuitry. Nonlinear digital cancellation may be employed to recreate and cancel these distortion products, but the associated complexity and power consumption must be considered. Local oscillator (LO) phase noise can pose problems as well. If a common LO is used for the transmitter and the receiver, the phase noise in the transmitted and the received SI will be completely correlated, enabling its cancellation in the receiver downmixer. However, delay in the SI channel will decorrelate the phase noise, resulting in residual SI that cannot be cancelled. Figure 1 depicts transceiver performance requirements calculated for two different SIC allocations across domains, one antenna-heavy and the other digital-heavy.

SI Channel Frequency Selectivity and Wideband RF/Analog SI Cancellation

The wireless SI channel can be extremely frequency selective. Compact antennas can be quite narrowband, and the front-end filters that are commonly used in today's radios even more so. The wireless SI channel also includes reflections off nearby objects, which will feature a delay that depends on the distance of the object from the radio. Performing wideband RF/analog cancellation requires recreating the wireless SI channel in the RF/analog domain. Conventional analog/RF cancellers feature a frequency-flat magnitude and phase response, and will therefore achieve cancellation only over a narrow bandwidth (BW). Wideband SIC at RF based on time-domain equalization (essentially an RF finite impulse response [FIR] filter) has been reported in [1] using discrete components. However, the integration of nanosecond-scale RF delay lines on an IC is a formidable (perhaps impossible) challenge, and therefore alternate wideband analog/RF SIC techniques are required.

Compact FD Antenna Interfaces

FD radios employing a pair of antennas, one for transmit and one for receive, experience a direct trade-off between form factor and transmit-receive isolation arising from the antenna spacing and design. Therefore, techniques that can maintain or even enhance transmit-receive isolation, possibly through embedded cancellation, while maintaining a compact form factor are highly desirable. Compact FD antenna interfaces are also more readily compatible with multiple-input multiple-output (MIMO) and diversity applications, and promote channel reciprocity, which is useful at the higher layers. For highly form-factor-constrained mobile applications, single-antenna FD is required, necessitating the use of circulators. Traditionally, circulators have been implemented

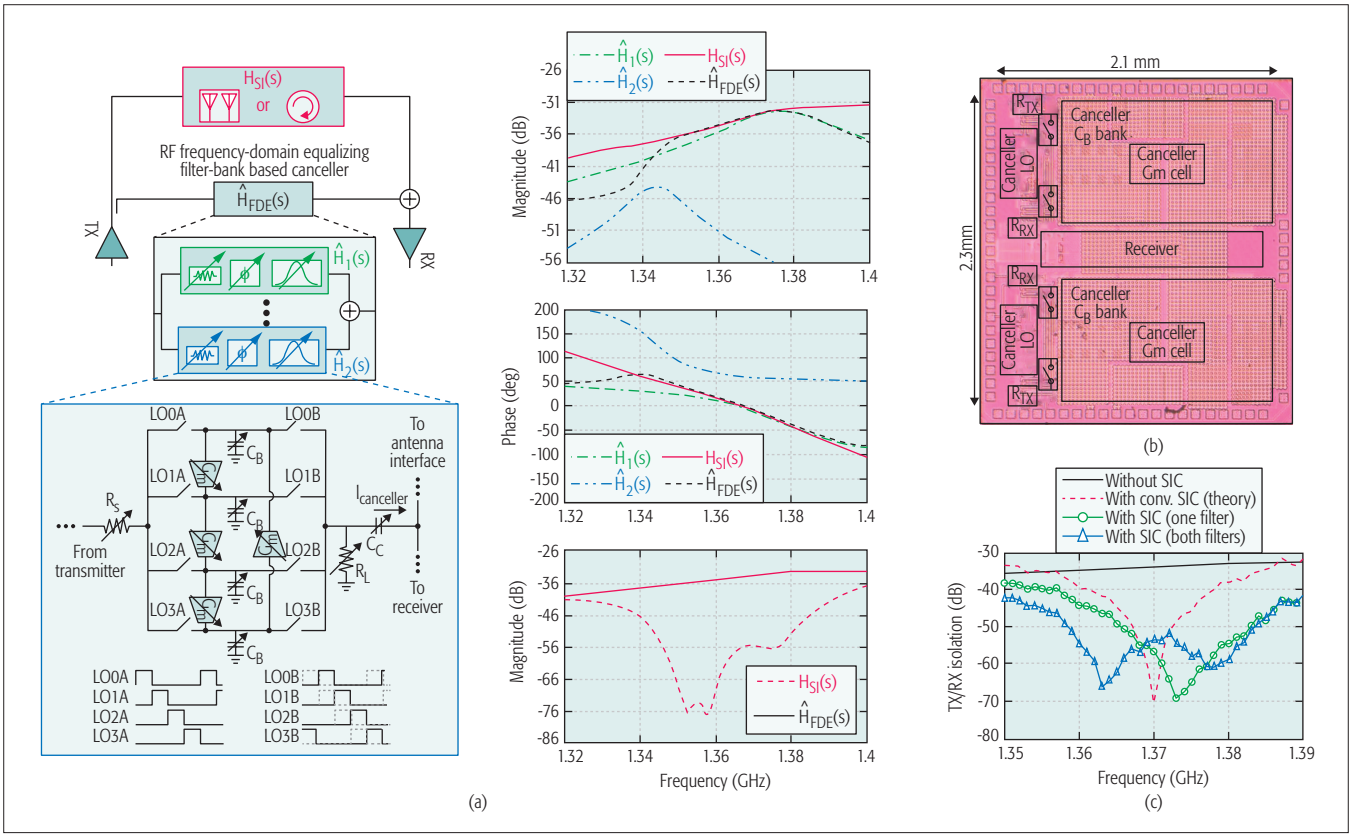


Figure 2. Integrated wideband RF SIC based on frequency-domain equalization (FDE): a) FDE concept and two-port Gm-C *N*-path filter with embedded variable attenuation and phase shift; b) chip photo of the implemented 0.8-to-1.4 GHz 65 nm CMOS FD receiver with FDE-based SIC in the RF domain featuring a bank of two filters; c) transmit-receive isolation of an antenna pair without SIC, with conventional SIC (theoretical), and with the proposed FDE-based SIC.

using ferrite materials, and are costly, bulky, and not compatible with IC technology. Novel techniques for high-performance non-magnetic integrated circulators are of high interest.

Adaptive Cancellation

The SIC in all domains must be reconfigurable and automatically adapt to changing operation conditions (e.g., supply voltage and temperature) and, most importantly, a changing electromagnetic (EM) environment (i.e., wireless SI channel), given the high level of cancellation required. This requires the periodic (or perhaps even continuous) usage of pilot signals to characterize the SI channel, the implementation of reconfigurable cancellers (which is more challenging in the antenna and RF domains), and the development of canceller adaptation algorithms.

Resource Allocation and Rate Gains for Networks with Integrated FD Radios, and Rethinking MAC Protocols

The benefits of enabling FD are clear: the uplink (UL) and downlink (DL) rates can theoretically be doubled (in both random access networks, e.g., WiFi, and small cell networks). That, of course, is true, provided that the SI is cancelled such that it becomes negligible at the receiver. Hence, most of the research on FD at the higher layers has focused on designing protocols and assessing the capacity gains while using models of recent laboratory bench-top FD implementations (e.g., [1]) and assuming perfect SI cancellation. However, given the special characteristics of IC-based SI cancellers, there is a need to understand the capacity gains and develop resource allocation algorithms while taking into account these characteristics. These algorithms will then serve as building blocks for the redesign of MAC protocols for FD networks with integrated FD radios.

Integrated RF Self-Interference Cancellation

To address the challenge related to integrated RF cancellation across a wide bandwidth (BW), we proposed a frequency-domain approach in contrast to the conventional time-domain delay-based RF FIR approach [3]. To enhance the cancellation BW, second-order reconfigurable bandpass filters (BPFs) with amplitude and phase control are introduced in the RF canceller (Fig. 2a). An RF canceller with a reconfigurable second-order RF BPF features four degrees of freedom (amplitude, phase, quality factor, and center frequency of the BPF). *This enables the replication of not just the amplitude and phase of the antenna interface isolation [$H_{SI}(s)$] at a frequency point, but also the slope of the amplitude and the slope of the phase (or group delay). The use of a bank of filters with independent BPF parameters enables such replication at multiple points in different sub-bands, further enhancing SIC BW. Essentially, this approach is frequency-domain equalization (FDE) in the RF domain.* In Fig. 2a, which represents a theoretical computation on the measured isolation of a

pair of 1.4 GHz antennas that are described in greater detail below, two BPFs with transfer functions $\hat{H}_1(s)$ and $\hat{H}_2(s)$ emulate the antenna interface isolation in two sub-bands, resulting in an order-of-magnitude improvement in the SIC BW over a conventional frequency-flat RF canceller.

For FDE, reconfigurable RF filters with very sharp frequency response (or high quality factor) are required. The achievable quality factor of conventional LC-based integrated RF filters has been limited by the quality factor of the inductors and capacitors that are available on silicon. However, recent research advances have revived a switched-capacitor circuit-design technique known as the *N*-path filter that enables the implementation of reconfigurable, high-quality filters at RF in nanoscale CMOS IC technology [10]. Figure 2a depicts a two-port *N*-path filter, where R_S and R_L are the resistive loads at the transmit and receive sides, respectively. C_C weakly couples the cancellation signal to the receiver input for SIC. The quality factor of an *N*-path filter may be reconfigured via the baseband capacitor C_B, given fixed R_S and R_L. Through clockwise/counter-clockwise (only counter-clockwise connection is shown in Fig. 2a for simplicity) connected reconfigurable transconductors (G_m), an upward/downward frequency offset with respect to the switching frequency can be introduced without having to change the clock frequency [10]. Variable attenuation can be introduced by reconfiguring R_S and R_L relative to each other. *Interestingly, phase shifts can be embedded in a two-port* N*-path filter by phase shifting the clocks driving the switches on the output side relative to the input-side clocks as shown in Fig. 2a [3].* All in all, the ability to integrate reconfigurable high-quality RF filters on chip using switches and capacitors uniquely enables synthesis of nanosecond-scale delays through FDE over time-domain equalization.

A 0.8–1.4 GHz FD receiver IC prototype with the FDE-based RF SI canceller was designed and fabricated in a conventional 65 nm CMOS technology (Fig. 2b) [3]. For the SIC measurement results shown in Fig. 2c, we used a 1.4 GHz narrowband antenna-pair interface with peak isolation magnitude of 32 dB, peak isolation group delay of 9 ns, and 3 dB of isolation magnitude variation over 1.36 to 1.38 GHz. The SI canceller achieves a 20 dB cancellation BW of 15/25 MHz (one/two filters) in Fig. 2c. When a conventional frequency-flat amplitude- and phase-based canceller is used, the SIC BW is about 3 MHz (> 8× lower). The 20 MHz bandwidth over which the cancellation is achieved allows our FD receiver IC to support many advanced wireless standards including small-cell LTE and WiFi.

Compact Full Duplex Antennas Employing Reconfigurable Polarization-Based SI Cancellation

The suppression of the SI within the antenna interface itself has significant advantages, as it relaxes the dynamic range requirements on the RF, analog, and digital blocks in the receiver chain as well as the RF/analog and digital SIC circuits. *While contemplating SI suppression at the antenna, it is important to keep in mind that FD antenna interfaces must also exhibit a compact form factor, preserve their radiation patterns, and maintain SI suppression in the presence of a changing EM environment*. Prior antenna-domain SI suppression approaches only partially satisfy these requirements [11]. In particular, they are typically static approaches that are unable to respond to a changing environment.

The antenna-electronics interface offers a unique opportunity to blend EM, RF, analog, and digital concepts to create "smart" antennas that achieve novel functionality. *In particular, the antenna domain offers another degree of freedom, wave polarization, apart from the conventional amplitude, phase, and frequency, which are the workhorses of the electronic domain*. Orthogonal polarizations can be exploited to enhance transmit-receive isolation, and, *more interestingly, embed reconfigurable SIC*.

Using this insight, we recently developed a wideband reconfigurable polarization-based antenna cancellation technique for FD. The technique is depicted in Fig. 3a and employs a pair of compact co-located antennas for the TX and the RX that use orthogonal polarizations to enhance the initial TX-to-RX isolation. Additionally, an auxiliary port that is co-polarized with the TX antenna is introduced on the RX antenna and terminated with a reconfigurable reflective termination (essentially a programmable filter). Since this port is co-polarized with the TX antenna, it "steals" a small portion of the transmitted signal, thus creating an indirect coupling path between the TX and RX ports. The signal in the indirect path is conditioned through the reflective termination and then couples to the RX port. By configuring the reflective termination appropriately, SIC can be achieved at the RX port. Through the implementation of a higher-order reflection termination, our technique can mimic the direct path's magnitude and phase as well as their slopes at multiple frequency points to achieve wideband cancellation in a manner similar to the FDE technique described earlier. The electronically programmable nature of this reflective termination also allows SIC to be reconfigured in-field in the face of a changing EM environment.

A 4.6 GHz antenna prototype employing this technique was built (Fig. 3b), and achieves more than 50 dB isolation over 300 MHz bandwidth [8]. This represents a 20× improvement in isolation bandwidth over conventional techniques. A strong reflector (a metallic plate) was also placed near the antenna during measurement, and the ability to reconfigure and recover the cancellation despite the reflector's presence was demonstrated (Fig. 3b).

Such antenna-electronics co-design techniques readily translate to higher frequencies where antennas are naturally smaller and interface more tightly with the IC. Recently, we reported a transceiver IC that combines FD with millimeter-waves [4]. Merging millimeter-waves with FD can potentially offer the dual benefits of wide bandwidths and improved spectral efficiency, a step toward delivering the tremendous increase in capacity demanded by emerging wireless standards. As shown in Fig. 3c, our 60 GHz FD transceiver IC employs the reconfigurable wideband polarization-based antenna cancellation discussed earlier. The reconfigurable reflective termination is implemented on the chip. To improve the SI suppression further, an RF canceller from the TX out-

The suppression of the SI within the antenna interface itself has significant advantages, as it relaxes the dynamic range requirements on the RF, analog, and digital blocks in the receiver chain as well as the RF/analog and digital SIC circuits.

put to the low-noise amplifier (LNA) output is also included in the transceiver. The complete 60 GHz FD transceiver architecture and its chip microphotograph are shown in Fig. 3d. It is implemented in a 45 nm silicon-on-insulator (SOI) CMOS process and achieves the highest integration level among FD transceivers irrespective of the operation frequency. In conjunction with digital SIC implemented in MATLAB after capturing the baseband (BB) signals using an Agilent 54855A oscilloscope (essentially an 8-bit 20 GSps ADC), a total SI suppression of nearly 80 dB was achieved over 1 GHz BW, enabling the world's first millimeter-wave FD link over a distance of almost 1 m. Figure 3e shows the demonstration setup using a 100 MHz offset continuous wave (CW) signal and

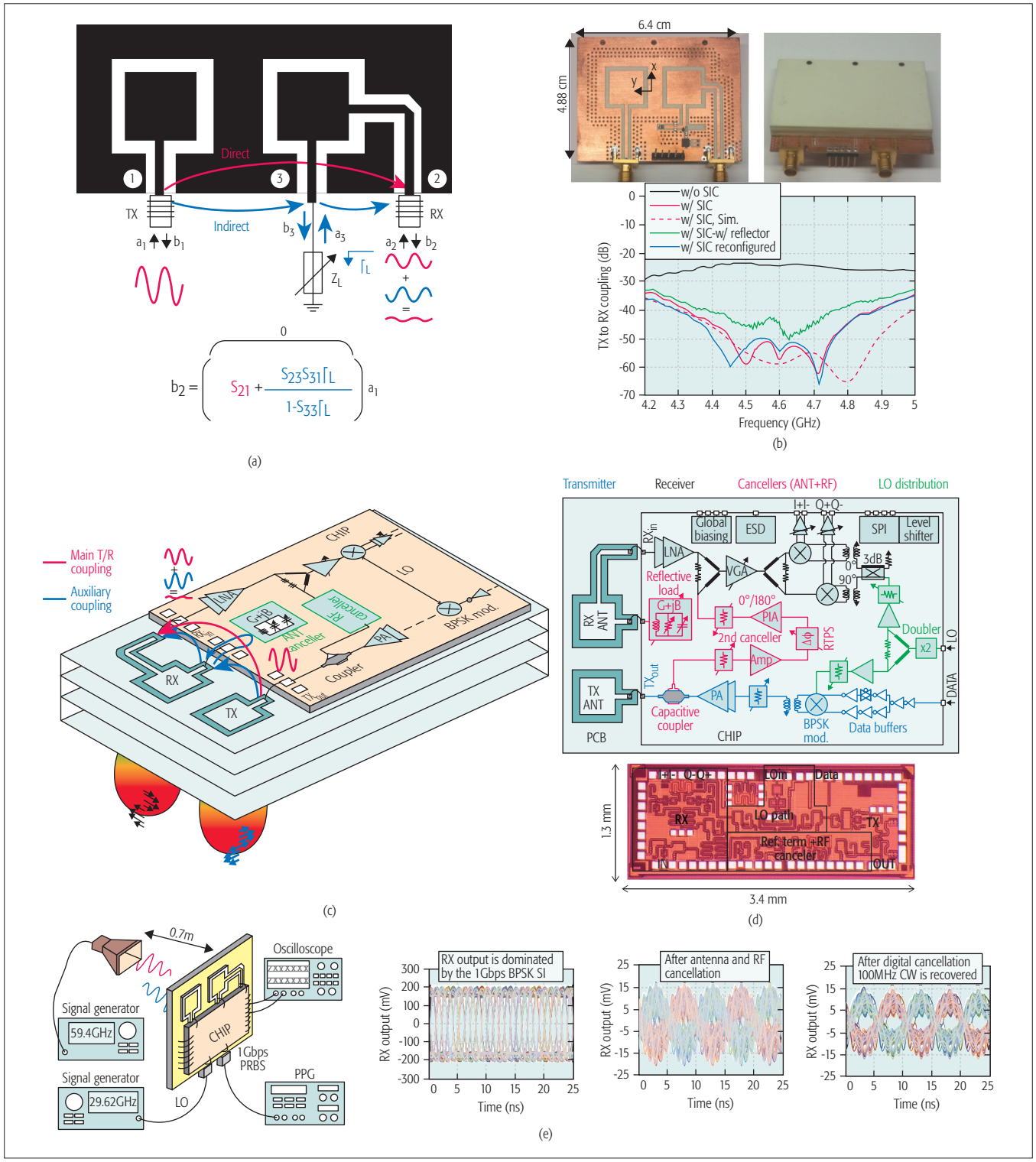


Figure 3. Polarization-based reconfigurable wideband antenna cancellation: a) concept; b) 4.6 GHz TX/RX antenna pair prototype and associated measurement results; c) 3D implementation view of a 60 GHz FD transceiver employing the proposed antenna cancellation; d) 60 GHz fully integrated FD transceiver architecture and IC microphotograph; e) 60 GHz FD link setup and demonstration.

1 Gb/s binary phase shift keying (BPSK) as the received and the transmitted signal (SI), respectively. In the absence of antenna and RF SIC, the RX output is dominated by SI. Engaging the antenna and RF SIC enables the discerning of the desired signal. Digital cancellation in MATLAB (a 100-tap adaptive LMS filter with a settling time of 16 μs) further suppresses the SI, resulting in an even cleaner received signal with a signal-to-interference-noise-and-distortion ratio (SINDR) of 7.2 dB.

Integrated Non-Magnetic Circulator for Single-Antenna Full Duplex

Highly-form-factor-constrained mobile applications, particularly at RF frequencies where the wavelength is considerably higher, demand single-antenna solutions. Single-antenna FD also ensures channel reciprocity and compatibility with antenna diversity and MIMO concepts. However, *conventional single-antenna FD interfaces, that is, non-reciprocal circulators, rely on ferrite materials and biasing magnets, and are consequently bulky, expensive and incompatible with silicon integration.* Reciprocal circuits, such as electrical-balance duplexers, have been considered, but are limited by the fundamental minimum 3 dB loss in both TX-antenna (ANT) and ANT-RX paths.

As mentioned earlier, non-reciprocity and circulation have conventionally been achieved using the magneto-optic Faraday effect in ferrite materials. However, it has recently been shown that violating time invariance within a linear, passive material with symmetric permittivity and permeability tensors can introduce non-reciprocal wave propagation, enabling the construction of non-magnetic circulators [12]. However, these initial efforts have resulted in designs that are either lossy, highly nonlinear, or comparable in size to the wavelength, and are fundamentally not amenable to silicon integration. *Recently, we introduced a new non-magnetic CMOS-compatible circulator concept based on the phase-non-reciprocal behavior of linear, periodically time-varying (LPTV) two-port* N*-path filters that utilize staggered clock signals at the input and output [9].*

N-path filters, described earlier in the context of FDE, are a class of LPTV networks where the signal is periodically commutated through a bank of linear, time-invariant (LTI) networks. We found that when the non-overlapping clocks driving the input and output switch sets of a two-port *N*-path filter are phase shifted with respect to each other, a nonreciprocal phase shift is produced for signals traveling in the forward and reverse directions as they see a different ordering of the commutating switches (Fig. 4a). The magnitude response remains reciprocal and low-loss, similar to traditional *N*-path filters. To create non-reciprocal wave propagation, an *N*-path filter with ± 90° phase shift is placed inside a $3\lambda/4$ transmission line loop (Fig. 4b). This results in satisfaction of the boundary condition in one direction (−270° phase shift from the loop added with −90° from the *N*-path filter) and suppression of wave propagation in the other direction (−270° + 90° = −180°), effectively producing unidirectional circulation. Additionally, a three-port circulator can be realized by placing ports anywhere along the loop as long as they maintain a $\lambda/4$ circumferential distance between them. Interestingly, maximum linearity with respect to the TX port is achieved if the RX port is placed adjacent to the *N*-path filter ($l = 0$), since the inherent TX-RX isolation suppresses the voltage swing on either side of the *N*-path filter, enhancing its linearity.

A prototype circulator based on these concepts operating over 610-850 MHz was implemented in a 65 nm CMOS process. Measurements reveal 1.7 dB loss in TX-ANT and ANT-RX transmission, and broadband isolation better than 15 dB between TX and RX (the narrowband isolation can be as high as 50 dB). The in-band ANT-RX IIP3 is +8.7 dBm while the in-band TX-ANT IIP3 is +27.5 dBm (OIP3 = +25.8 dBm), two orders of magnitude higher due to the suppression of swing across the *N*-path filter. The measured clock feedthrough to the ANT port is −57 dBm, and IQ image rejection for TX-ANT transmission is 49 dB. Techniques such as device stacking in SOI CMOS can be explored to further enhance the TX-ANT linearity to meet the stringent requirements of commercial wireless standards. Clock feedthrough and IQ mismatch can be calibrated by sensing and injecting appropriate BB signals through the *N*-path filter capacitor nodes as shown previously in the literature.

A 610–850 MHz FD receiver IC prototype incorporating the non-magnetic *N*-path-filter-based passive circulator and additional analog BB SI cancellation (shown in Figs. 4c and 4d) was also designed and fabricated in the 65 nm CMOS process [5]. SI suppression of 42 dB is achieved across the circulator and analog BB SIC over a BW of 12 MHz. Digital SIC has also been implemented in MATLAB after capturing the BB signals using an oscilloscope (effectively an 8-bit 40 MS/s ADC) (Fig. 4e). The digital SIC cancels not only the main SI but also the IM3 distortion generated on the SI by the circulator, receiver, and canceller. A total of 164 canceller coefficients are trained by 800 sample points. After digital SIC, the main SI tones are at the −92 dBm noise floor, while the SI IM3 tones are 8 dB below for −7 dBm TX average power. This corresponds to an overall SI suppression of 85 dB for the FD receiver.

Resource Allocation and Achievable Rate Improvements in FD

Cancelling SI to a negligible level is extremely challenging. Therefore, we have been considering the following questions: *How much can be gained given a realistic canceller residual SI profile and signal strength? And when does it make sense to use FD over legacy time-division duplex (TDD)?* These questions need to be addressed in the context of a hybrid network with both FD and legacy half-duplex (HD) nodes, as illustrated in Fig. 5a. As a first step and in order to obtain fundamental understanding, in [13, 14] we focused on a single FD bidirectional link with orthogonal channels (e.g., orthogonal frequency division multiplexing — OFDM), and obtained analytic and algorithmic resource (power, time, and frequency) allocation results that quantify the achievable rate improvements as a function of SI-to-noise ratios (XINRs) and signal-to-noise ratios (SNRs).

To model achievable rates on the UL and DL, we used Shannon's capacity formula. We

Conventional single-antenna FD interfaces, that is, non-reciprocal circulators, rely on ferrite materials and biasing magnets, and are consequently bulky, expensive, and incompatible with silicon integration. Reciprocal circuits, such as electrical-balance duplexers, have been considered, but are limited by the fundamental minimum 3 dB loss in both TX-antenna (ANT) and ANT-RX paths.

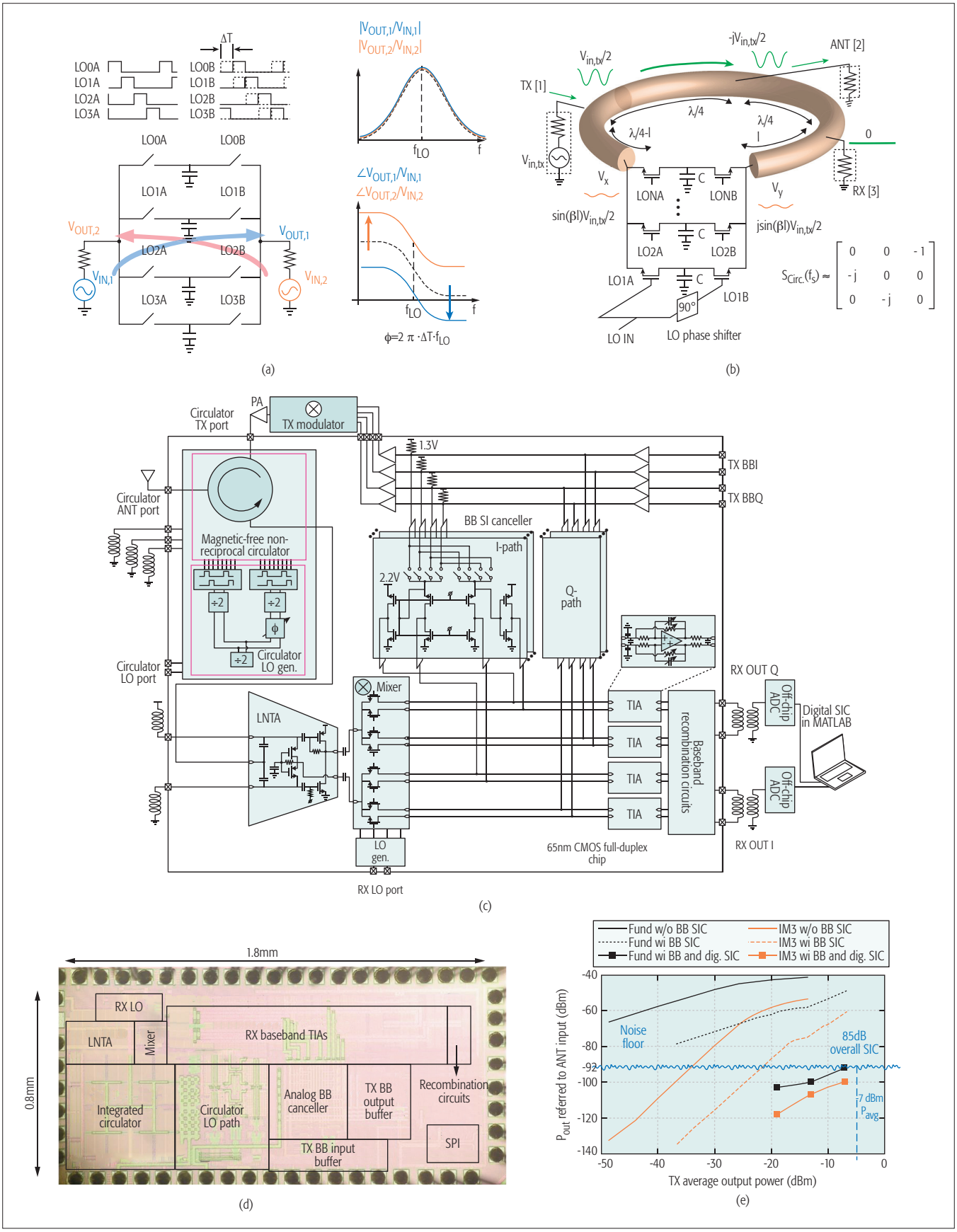


Figure 4. Integrated non-magnetic circulator for single-antenna FD: a) non-reciprocity induced by phase-shifted *N*-path commutation; b) 3-port circulator structure obtained by placing the non-reciprocal two-port *N*-path filter with ± 90° phase shift within a 3λ/4 transmission line loop; c) block diagram and schematic of the implemented 65 nm CMOS FD receiver with non-magnetic circulator and additional analog BB SI cancellation; d) chip microphotograph of the implemented FD receiver IC; e) measured two-tone SI test with SI suppression across circulator, analog BB and digital domains.

assumed that the residual SI is a constant fraction of the transmitted signal, where the "constant fraction" can be different for different channels. Under such a model, assuming that there are K orthogonal channels, the rate on the DL can be written as

$$r_b = \sum_{k=1}^{K} \log\left(1 + \frac{\mathrm{SNR}_{\mathrm{bm},k}}{1 + \mathrm{XINR}_{\mathrm{mm},k}}\right),$$

where $\mathrm{SNR}_{\mathrm{bm},k}$ is the SNR at the mobile station (MS) on channel k and $\mathrm{XINR}_{\mathrm{mm},k}$ is the XINR at the MS on channel k. Similarly, the UL rate is

$$r_m = \sum_{k=1}^{K} \log\left(1 + \frac{\mathrm{SNR}_{\mathrm{mb},k}}{1 + \mathrm{XINR}_{\mathrm{bb},k}}\right).$$

Note that the noise levels at the MS and the BS and over orthogonal channels are not assumed to be equal.

We now briefly outline a few main results. As illustrated in Fig. 5b, for a given FD rate pair (r_b, r_m), we defined the *FD rate improvement* p as the (positive) number for which $(r_b/p, r_m/p)$ is at the boundary of the corresponding TDD capacity region.

First, we focused on maximizing the sum of the UL and DL rates (referred to as the *sum rate*) when *only a single channel is considered*. We showed that *if any FD rate pair has higher sum than the maximum TDD rate, the maximum sum rate is obtained when both UL and DL TX power levels are set to their maximum values*. Let (s_b, s_m) denote the rate pair corresponding to the maximum TX power levels. Figures 5c and 5d show the rate improvements at (s_b, s_m) when the XINR at the BS is 0 dB, as a function of the XINR at the MS and SNRs at the UL and DL. As Figs. 5c and 5d suggest, *to obtain non-negligible rate improvements from FD, SNRs need to be sufficiently high compared to the XINRs*.

For a *general number of channels*, the problem of allocating power levels to orthogonal UL and DL channels such that the sum rate is maximized is non-convex. However, we showed that under mild restrictions, the problem is, in fact, amenable to efficient optimization methods. The restrictions impose a lower bound on the amount of SIC that needs to be obtained for given levels of SNR. We show that if these restrictions are not satisfied, FD cannot provide appreciable rate improvements. Based on realistic models of residual SI at the BS [1] and at the MS [3, 15], we demonstrated in [13] that *simple power allocation methods, such as the high SINR approximation power allocation, are near-optimal whenever the gains from FD are non-negligible*.

In [14] we focused on determining the *capacity region* of an FD link. This is equivalent to the problem of maximizing one of the (UL and DL) rates when the other is fixed. An FD capacity region is not convex in general. However, the region can be "convexified" through time sharing between different FD rate pairs. We refer to a convexified capacity region as the time-division full-duplex (TDFD) capacity region. A TDFD capacity region generally provides higher rates than its corresponding FD region. Moreover, *a convex (TDFD) capacity region is desirable, since most resource allocation and scheduling algorithms rely on convexity, as providing performance guarantees for a non-convex region is hard*.

Although the problem of determining either the FD or the TDFD capacity region is non-convex in general, we developed an algorithm (AltMax) that determines the TDFD region under mild restrictions and is guaranteed to converge to a stationary point, which in practice is a global optimum. The restrictions lead to suboptimality mainly in the region where, on average, XINR over channels is high compared to the average SNR over channels. Building on insights from our numerical experiments, we also developed a simple heuristic with similar performance but lower running time. The rate improvements obtained by AltMax and the heuristic are illustrated in Figs. 5e and 5f for the residual SI of the FD receiver implemented in [15], and Figs. 5g and 5h for the residual SI of the FD receiver implemented in [3]. In the figures, we assume that 110 dB is required for the TX signal to be cancelled to the noise level, as in, for example, [1]. We can observe from Figs. 5e–5h that as the average SNR increases, the rate improvements increase.[1]

Comparing the rate improvements for the canceller from [16] with the rate improvements for the canceller from [3] (i.e., comparing Fig. 5e with Fig. 5g and Fig. 5f with Fig. 5h, it is not difficult to observe, as expected, that more broadband cancellation ([3] vs. [15]) provides higher rate improvements. Moreover, we can observe that in the regions of higher SNRs, where the rate improvements are higher (Figs. 5f and 5h), AltMax and the heuristic provide almost indistinguishable solutions. Therefore, *in the regions of high rate improvements, the TDFD capacity region can be determined near optimally with a simple heuristic*.

Finally, we note that FD has other advantages in addition to potential 2× rate improvement. For example, in WiFi systems, FD can reduce collisions and, consequently, reduce the packet loss.

Conclusion

In summary, the Columbia FlexICoN project has been focusing on IC-based FD transceivers spanning RF to millimeter-wave, reconfigurable antenna cancellation, non-magnetic CMOS circulators, and MAC layer algorithms based on realistic hardware models. While exciting progress has been made in the last few years by the research community as a whole, several problems remain to be solved before FD wireless can become a widely deployed reality. Continued improvements are necessary in IC-based FD transceivers toward increased total cancellation over wide BWs and support for higher TX power levels through improved RX and circulator linearity. Incorporation of FD in large-scale phased-array transceivers is another open research problem. The extension of IC-based SIC concepts to MIMO transceivers is an important and formidable challenge. In a MIMO transceiver, SI will exist between every TX-RX pair, and a brute force implementation will cause canceller complexity to scale quadratically with the number of MIMO elements. At the higher layers, while extensive recent research has been devoted to this area, the implications of different possible PHY layer implementations are still not fully understood. Moreover, there are still several important open problems related to MAC layer design for both cellular and random access networks. Specifically, it is necessary to design algorithms that support asymmetric UL and DL traffic requirements and provide fairness

While exciting progress has been made in the last few years by the research community as a whole, several problems remain to be solved before FD wireless can become a widely-deployed reality. Continued improvements are necessary in IC-based FD transceivers towards increased total cancellation over wide BWs and support for higher TX power levels through improved RX and circulator linearity.

[1] In fact, it can be shown that for any finite values of XINR over channels, if it was possible to increase the SNR to infinity, the rate improvements would reach the theoretical upper bound of 2.

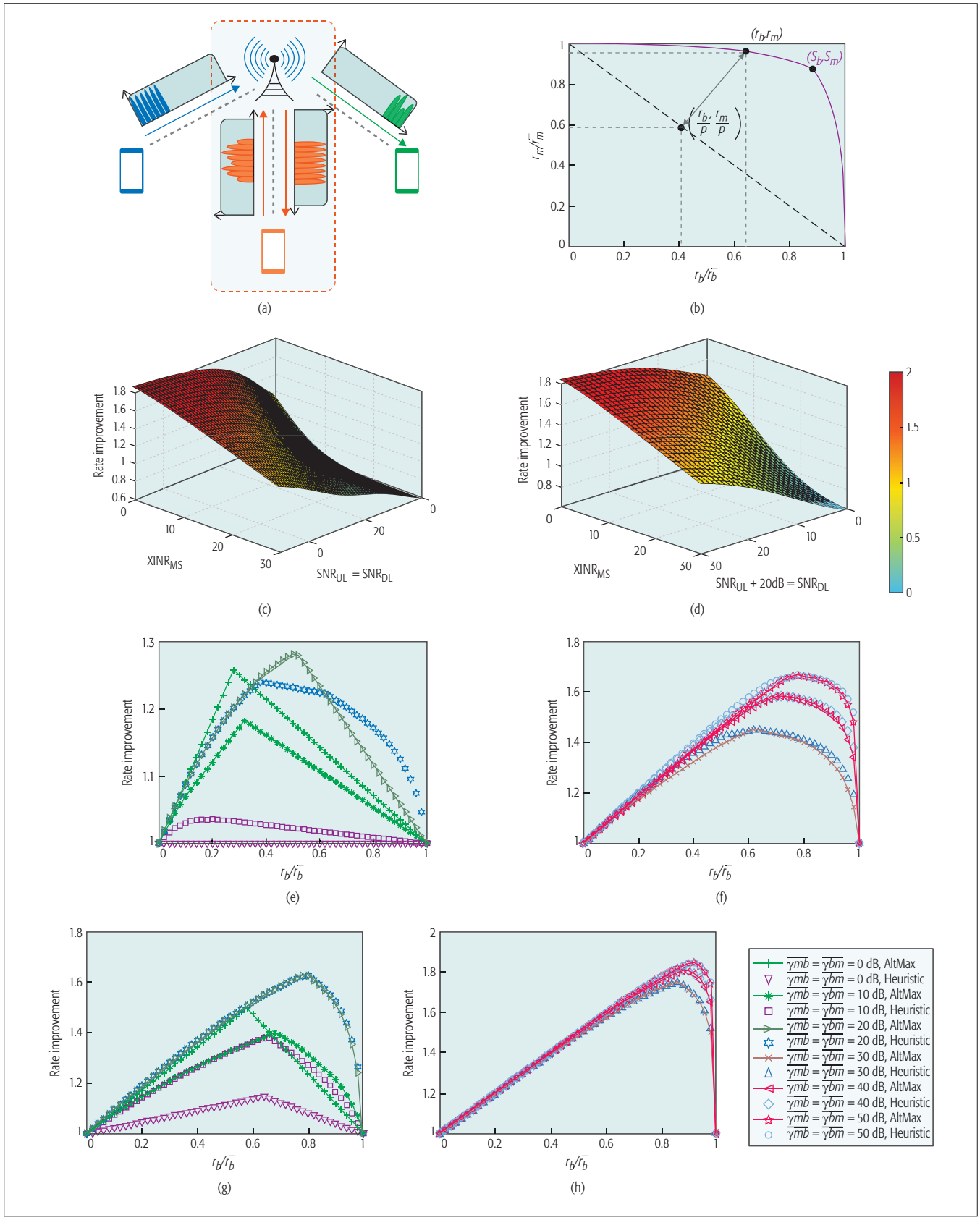


Figure 5. Resource allocation and rate gains in FD systems: a) illustration of a network with FD and HD users, with the FD link outlined by a box; b) definition of rate improvement; c) rate improvement at (s_b, s_m) for equal SNRs on UL and DL; d) rate improvement at (s_b, s_m) for 20 dB higher SNR at the DL than at the UL; e) rate improvements for fixed r_b and maximized r_m, for the FD receiver from [15] and average SNR (denoted as $\gamma_{mb} = \gamma_{bm}$) up to 20 dB; f) rate improvements for fixed r_b and maximized r_m, for the FD receiver from [15] and average SNR from 30 dB to 50 dB; g) rate improvements for fixed r_b and maximized r_m, for the FD receiver from [3] and average SNR up to 20 dB; h) rate improvements for fixed r_b and maximized r_m, for the FD receiver from [3] and average SNR from 30 dB to 50 dB.

in networks composed of both FD and legacy HD nodes. In addition, it is important to consider interference management in OFDM networks jointly at the PHY and MAC layers.

Integrated FD is suitable for many applications ranging from backhaul, relays, and WiFi to small-cell LTE (where cancellation requirements are relaxed compared to macrocells). Although antenna cancellation techniques provide wider BW and relax dynamic range requirements, they tend to have large form factors, making them more practical for point-to-point/less form-factor-constrained applications like backhaul and relays. On the other hand, RF cancellation and on-chip circulator-based FD works generally result in smaller form factors, making them more feasible for mobile devices for WiFi and small cells with future improvements in cancellation BW and power handling capability.

Acknowledgments

This work was supported in part by the DARPA RF-FPGA program, the DARPA ACT program, NSF grant ECCS-1547406, the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no. [PIIF-GA-2013-629740].11, and a Qualcomm Innovation Fellowship.

References

[1] D. Bharadia, E. McMilin, and S. Katti, "Full Duplex Radios," *Proc. ACM SIGCOMM '13*, Oct. 2013, pp. 375–86.
[2] S. Enserink *et al.*, "Joint Analog and Digital Interference Cancellation," *Proc. 2014 IEEE MTT-S Int'l. Microwave Symp.*, June 2014, pp. 1–3.
[3] J. Zhou *et al.*, "Integrated Wideband Self-Interference Cancellation in the RF Domain for FDD and Full-Duplex Wireless," *IEEE J. Solid-State Circuits*, vol. 50, no. 12, Dec. 2015, pp. 3015–31.
[4] T. Dinc, A. Chakrabarti, and H. Krishnaswamy, "A 60GHz CMOS Full-Duplex Transceiver and Link with Polarization-Based Antenna and RF Cancellation," *IEEE J. Solid-State Circuits*, vol. 51, no. 5, 2016, pp. 1125–40.
[5] J. Zhou, N. Reiskarimian, and H. Krishnaswamy, "Receiver with Integrated Magnetic-Free N-Path-Filter-Based Non-Reciprocal Circulator and Baseband Self-Interference Cancellation for Full-Duplex Wireless," *Proc. IEEE ISSCC '16*, Feb. 2016, pp. 178–80.
[6] D. Yang, H. Yuksel, and A. Molnar, "A Wideband Highly Integrated and Widely Tunable Transceiver for In-Band Full-Duplex Communication," *IEEE J. Solid-State Circuits*, vol. 50, no. 5, May 2015, pp. 1189–1202.
[7] D. J. van den Broek, E. A. M. Klumperink, and B. Nauta, "An in-Band Full-Duplex Radio Receiver with a Passive Vector Modulator Downmixer for Self-Interference Cancellation," *IEEE J. Solid-State Circuits*, vol. 50, no. 12, Dec. 2015, pp. 3003–14.
[8] T. Dinc and H. Krishnaswamy, "A T/R Antenna Pair with Polarization-Based Wideband Reconfigurable Self-Interference Cancellation for Simultaneous Transmit and Receive," *Proc. IEEE IMS'15*, 2015, pp. 1–4.
[9] N. Reiskarimian and H. Krishnaswamy, "Magnetic-Free Non-Reciprocity Based on Staggered Commutation," *Nature Commun.*, vol. 7, no. 4, Apr. 2016.
[10] M. Darvishi *et al.*, "Widely Tunable 4th Order Switched Gm-C Band-Pass Filter Based on N-Path Filters," *IEEE J. Solid-State Circuits*, vol. 47, no. 12, Dec 2012, pp. 3105–19.
[11] E. Yetisir, C.-C. Chen, and J. Volakis, "Low-Profile UWB 2-Port Antenna With High Isolation," *IEEE Antennas Wireless Propag. Lett.*, vol. 13, 2014, pp. 55–58.
[12] N. A. Estep, D. L. Sounas, and A. Alu, "Magnetless Microwave Circulators Based on Spatiotemporally Modulated Rings of Coupled Resonators," *IEEE Trans. Microwave Theory Tech.*, vol. 64, no. 2, Feb. 2016, pp. 502–18.
[13] J. Marasevic *et al.*, "Resource Allocation and Rate Gains in Practical Full-Duplex Systems," *IEEE/ACM Trans. Net.*, vol. 25, no. 1, Feb. 2017, pp. 292–305.
[14] J. Marasevic and G. Zussman, "On the Capacity Regions of Single-Channel and Multi-Channel Full-Duplex Links," *Proc. ACM MobiHoc '16*, 2016.
[15] J. Zhou *et al.*, "Low-Noise Active Cancellation of Transmitter Leakage and Transmitter Noise in Broadband Wireless Receivers for FDD/Co-Existence," *IEEE J. Solid-State Circuits*, vol. 49, no. 12, Dec. 2014, pp. 3046–62.

Biographies

Jin Zhou [S'11, M'17] received his B.S. degree in electronics science and technology from Wuhan University, China, in 2008, his M.S. degree in microelectronics from Fudan University, Shanghai, China, in 2011, and his Ph.D. degree in electrical engineering from Columbia University, New York, in 2017. From 2011 to 2012, he worked as an RF integrated circuits design engineer with MediaTek Singapore. In 2017, he joined the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign as an assistant professor.

Negar Reiskarimian [S'15] received her B.Sc. and M.Sc. degrees in telecommunication systems and microelectronic circuits, both from Sharif University of Technology, Tehran, Iran, in 2011 and 2013. She is currently pursuing her Ph.D. degree in electrical engineering at Columbia University. She is the recipient of a 2016 Qualcomm Innovation Fellowship and the 2017 IEEE SSCS Predoctoral Achievement Award, the Analog Devices Outstanding Student Designer Award, and an IEEE MTT-S Graduate Fellowship.

Jelena Diakonikolas received her B.Sc. degree from the University of Belgrade in 2011, and her M.S. and Ph.D. degrees from Columbia University in 2012 and 2016. She is currently a postdoctoral associate at Boston University. Her research focuses on principled design of networked systems. She is a recipient of the M.S. Award of Excellence and the Jacob Millman Prize for Excellence in Teaching Assistance from Columbia University. She is a co-winner of the Qualcomm Innovation Fellowship 2015 award, and was listed as one of the "10 Women in Networking/Communications That You Should Watch" in 2016.

Tolga Dinc [S'09] received his B.S. and M.S. degrees in electrical engineering from Sabanci University, Istanbul, Turkey, in 2010 and 2012, respectively. He is currently working toward a Ph.D. degree in electrical engineering at Columbia University. He is a recipient of several honors and awards including the IEEE RFIC Symposium Best Student Paper Award (1st Place) in 2015.

Tingjun Chen [S'16] received his B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2014. and his M.S. degree in electrical engineering from Columbia University in 2015. He is currently a Ph.D. student in electrical engineering at Columbia University. His research interests are in algorithms, optimization, and system design in the Internet of Things, energy harvesting networks, full-duplex networks, and 5G networks. He received the Wei Family Private Foundation Fellowship, the Columbia University Electrical Engineering Armstrong Memorial Award, and the ACM CoNEXT 2016 Best Paper Award.

Gil Zussman [S'02, M'05, SM'07] received his Ph.D. degree in electrical engineering from the Technion in 2004 and was a postdoctoral associate at MIT in 2004–2007. Since 2007 he has been with Columbia University where he is currently an associate professor of electrical engineering. His research interests are in the areas of wireless, mobile, and resilient networks. He received the Fulbright Fellowship, DTRA Young Investigator Award, NSF CAREER Award, and two Marie Curie Fellowships. He was the PI of a team that won 1st place in the 2009 Vodafone Americas Foundation Wireless Innovation Project competition. He is a co-recipient of seven best paper awards, including the ACM SIGMETRICS/IFIP Performance '06 Best Paper Award, the 2011 IEEE Communications Society Award for Advances in Communication, and the ACM CoNEXT'16 Best Paper Award.

Harish Krishnaswamy [S'03, M'09] received his B.Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, India, in 2001, and his M.S. and Ph.D. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, in 2003 and 2009, respectively. In 2009, he joined the Electrical Engineering Department, Columbia University, where he is currently an associate professor. His research interests broadly span integrated devices, circuits, and systems for a variety of RF, mmWave and sub-mmWave applications. He serves as a member of the TPC of several conferences. He is currently serving as an IEEE SSCS Distinguished Lecturer for 2017 and 2018. He was the recipient of the IEEE International Solid-State Circuits Conference Lewis Winner Award for Outstanding Paper in 2007, the Best Thesis in Experimental Research Award from the USC Viterbi School of Engineering in 2009, the DARPA Young Faculty Award in 2011, a 2014 IBM Faculty Award, and the 2015 IEEE RFIC Symposium Best Student Paper Award, 1st Place.

It is necessary to design algorithms that support asymmetric UL and DL traffic requirements and provide fairness in networks composed of both FD and legacy HD nodes. In addition, it is important to consider interference management in OFDM networks jointly at the PHY and MAC layers.

ACCEPTED FROM OPEN CALL

3D Channel Models: Principles, Characteristics, and System Implications

Reham Nemer Almesaeed, Araz Sabir Ameen, Evangelos Mellios, Angela Doufexi, and Andrew Nix

The authors present a comprehensive review of the principles and characteristics of 3D channel models. They propose a framework for a 3D channel extension of the widely used 2D 3GPP/ITU generic channel model. They describe the main components and challenges of the newly proposed 3D channel model and the motivations that lie behind them.

ABSTRACT

This article presents a comprehensive review of the principles and characteristics of 3D channel models. We propose a framework for a 3D channel extension of the widely used 2D 3GPP/ITU generic channel model. We describe the main components and challenges of the newly proposed 3D channel model and the motivations that lie behind them. 3D channel models specify multipath elevation angles as well as azimuth (or horizontal plane) angles. This enables the evaluation of 3D MIMO techniques such as FD MIMO and per user 3D beamforming. We also provide a state-of-the-art review on the evolution of channel models. The article ends with a discussion on the impact of 3D channel modeling on system-level performance.

INTRODUCTION

Most of the directional models in the literature, including the standardized ones, concentrate on the direction data in the azimuth-only plane. However, many measurement campaigns have demonstrated that elevation angles have a significant impact on communication system-level performance [1]. The development of a 3D channel model opened up possibilities for a variety of strategies including full dimension multiple-input multiple-output (FD-MIMO), user-specific elevation beamforming, and cell splitting by benefiting from the richness of the real wireless channel [2]. In conventional MIMO systems, the antenna elements are deployed linearly in the azimuth plane only. However, when the linear antenna elements are extended to the 2D plane, the elevation angle should also be considered to obtain accurate correlation measure of the MIMO system. This requires both the azimuth and elevation angles to be represented in the propagation model.

More analysis is required in terms of modeling the elevation related statistics such as the elevation spread and distribution, and the power elevation spectrum. Research into the elevation angle domain can be traced back to 1970 by T. Aulin, who extended Clark's scattering model to the 3D domain [3]. In this work, the author assumed a rectangular and truncated cosine function for the elevation power angular spectrum (PAS) for the elevation angles. Elevation studies fall into two categories:

- Modeling of elevation statistics at the user equipment (UE)
- Modeling of elevation at the BS

However, most of the studies in the literature have focused on elevation spectra at the UE, as they have greater spread compared to the base station (BS)

UE measurements showed that multi path components (MPCs) arriving at the UE via over-the-rooftop propagation tend to have higher elevation angle spreads compared to MPCs that wave-guided in street canyons [4]. Extensive measurements have been carried out in [5], which showed that a double exponential function is a good approximation for the elevation power spectrum measured at the UE, with elevation spreads typically in the range of 10–15°. At the BS, elevation spread is considerably smaller according to [6]. In these studies the authors concluded that over-the-rooftop propagation and wave guiding in the street provide different contributions to the elevation spectrum, where clustering was also observed. In addition, measurement campaigns were carried out, such as in [7], to investigate the 3D characteristics of the wireless channel and to propose elevation statistics of mainly urban environments in macro and microcells. These studies assumed different antenna configurations, and also different propagation conditions such as indoor-to-outdoor propagation. Furthermore, a high resolution 3D model of the direction of arrival of the MPCs in an urban environment was proposed in [1], and the elevation dependence of the impinging power was investigated. Further analysis of the impact of the 3D component on the antenna correlation and gain imbalance in MIMO systems was presented in [8].

The motivation behind 3D channel modeling encouraged standardization bodies such as the Third Gneration Partnership Project (3GPP) to work on defining future mobile communication standards that provide accurate 3D channel models and help in evaluating the potential of advanced MIMO techniques [9]. Topics of particular interest are the dependence of the mean elevation and elevation spread on distance, the impact of BS height on elevation spread, and the best modeling of elevation statistical distributions. For example, the WINNER+ model has reported results in terms of the angle spread (AS) at both the UE and BS in an attempt to extend from a 2D into a 3D model [10]. In January 2013, 3GPP

Digital Object Identifier:
10.1109/MCOM.2017.1500505

Reham Nemer Almesaeed, Evangelos Mellios, Angela Doufexi, and Andrew Nix are with the University of Bristol; Araz Sabir Ameen is with the University of Sulaimani.

Working Group 1 (WG1) launched discussions on a 3D channel model, and since then many proposals have been reported considering different propagation environments [11]. The 3GPP 3D channel model provides more flexibility for the elevation dimension, which allows better modeling of the 2D antenna system. The contribution of this article is to highlight the importance of elevation angle modeling and its impact on system-level performance. It also reviews the authors' proposed model in this context, which is based on extensive ray tracing measurements to give more accurate modeling of the 3D channel compared to the models based on measurement campaigns. The latter is limited by the number of measurements and the deployed antenna type. Unlike the WINNER+ model, which assumes fixed elevation angle spread per environment, the article shows the distance dependence of elevation spread. The authors' model of elevation spread is considered and partially implemented in the 3GPP 3D channel model [11], and the proposed model is shown to match well with the smaller number of measurement observation available at the time.

This article is organized as follows. We first present the characteristics and principles of 3D models and discuss the modeling challenges associated with 3D multipath. Next, we discuss the changes required to the existing 3GPP/International Telecommunication Union (ITU) 2D channel model to develop 3D channel realizations suitable for system-level studies. The impact of 3D multipath on system performance is addressed in the final section.

Characteristics of Elevation Angle

Analysis of Elevation Spread

3D channel models have been developed in several high-profile projects such as WINNER+ [10]. However, most of these models depend mainly on literature surveys rather than real-world measurements or ray tracing predictions. WINNER's model for the elevation spread follows the azimuth plane, where fixed elevation spread is assumed per environment. For example, in urban macro environments the average elevation spread is 3° despite the UE's location from the BS. However, based on deterministic predictions from our validated 3D ray tracing engine [12], we have observed that the mean elevation angle spread decreases as a function of distance from the BS.

The tracer engine provides information on the power, phase, propagation time, angle of arrival, (AOA) and angle of departure (AOD) in both elevation and azimuth of each of the multipath components linking the BS to the UE. Ray geometries are computed using site-specific geographic databases for terrain, buildings, and foliage. Figure 1 presents an intuitive explanation of the distance dependency of the elevation spread. Figure 1 considers a simple geometric model for the elevation dimension. As shown in the figure, as the UE separation distance increases from the BS, the elevation angle spread decreases at the UE. The same trends were observed in the ray tracing data. Figure 2 shows the range of predicted angular spreads for different UE locations from the BS generated from the ray tracer predictions in central Bristol, United Kingdom. Figures 2a and 2b show the mean and variance of the log of the elevation spread at the UE as a function of separation distance from the BS. The mean elevation spread can be seen to decrease as a function of one-over-distance for this urban macro environment in non-line-of-sight (NLOS) conditions. According to Fig. 2, we also observe that the range of elevation angular spreads depends on the heights of the BS and UE. A comparison of the complementary distribution function of the elevation arrival and departure angular spread is shown in Figs. 2c and 2d for a large number of UEs distributed in the distance range of 50–1000 m from the BS. It is clearly shown that as the BS height increases, the range of angular spreads increases in the elevation plane. Please note that the term H10 in the figure refers to the BS height of 10 m and so on.

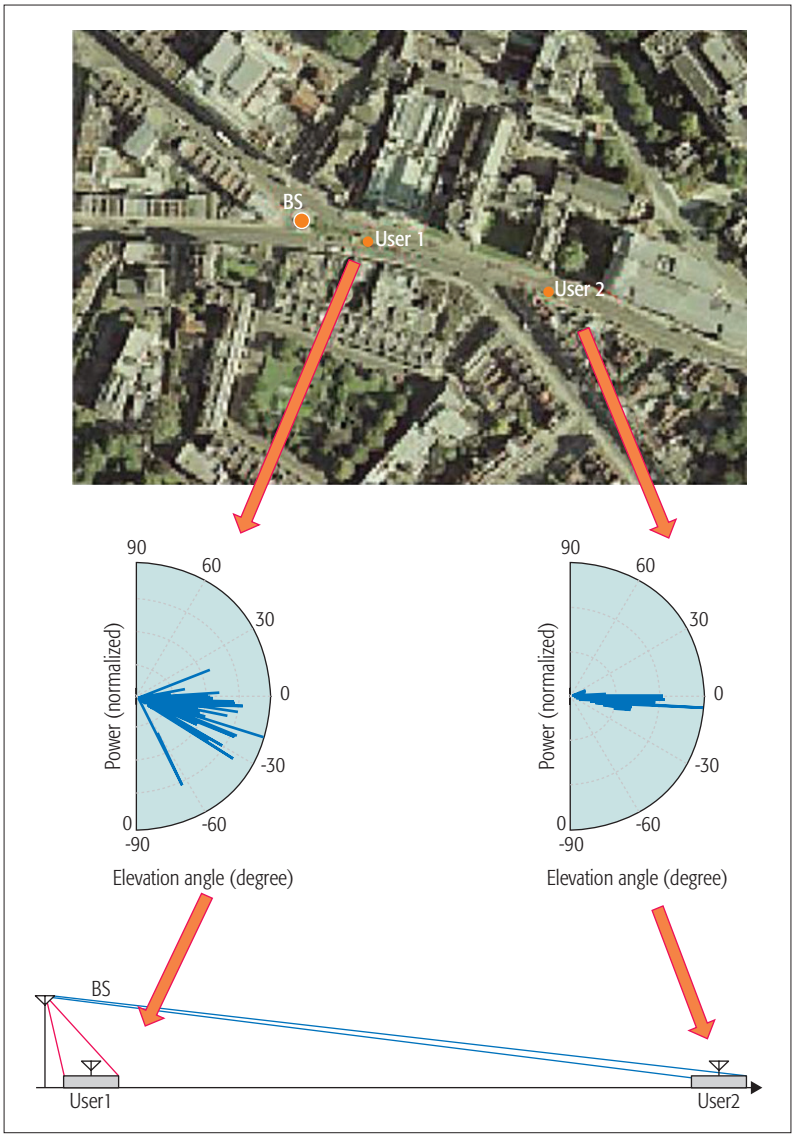


Figure 1. Intuitive explanation of distance dependent elevation angle spread.

Open Issues in Elevation Domain Modeling

The introduction of 3D channel modeling requires current 2D models to extend their propagation statistics to address the elevation domain. This subsection discusses some of the issues encountered when extending a 2D geometric-based sto-

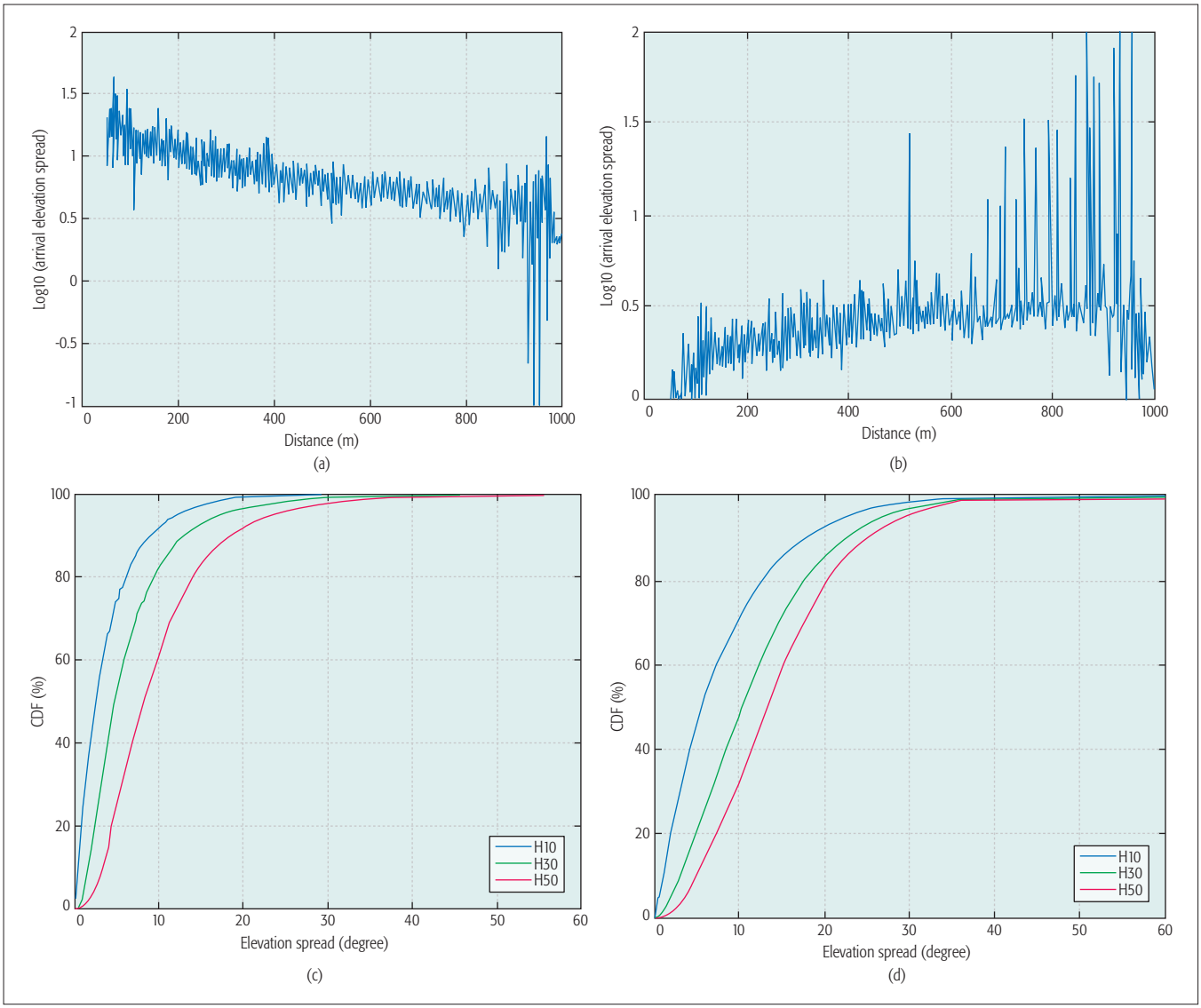


Figure 2. Analysis of predicted elevation angular spread: mean of arrival elevation spread vs. distance; b) variance of arrival elevation vs. distance; c) arrival elevation spread for different BS heights in LOS; d) .arrival elevation spread for different BS heights in NLOS.

chastic channel model to consider multipath in the elevation and azimuth domains.

Azimuth and elevation angular dependence: Since the elevation and azimuth angles of electromagnetic waves depend on the position of scatterers and reflectors in 3D space, the modeling of elevation angular information should be correlated with the azimuth plane. This is not taken into consideration in standardized channel models such as WINNER+, which proposes a set of elevation statistics (e.g., spread, PAS) that are independent of the azimuth statistics.

Elevation angular spread and spatial correlation: The power elevation spectrum (PES) is another important statistical property of the wireless channel and plays an important role in determining the spatial correlation present in MIMO systems. Most of the current models (e.g., the WINNER+ model) use the lognormal distribution to fit the azimuth angular spread (AS) for both BS and UE. Enhanced mathematical models can be explored to better fit the elevation statistics.

Antenna Radiation and Cross Polar Discrimination (XPD): The incorporation of elevation angles in the generation of the channel response requires complete 3D antenna patterns. In the case of a polarized MIMO system, the XPD ratio plays an important factor in determining the dependence of the antenna patterns at different polarizations. This factor depends on the angular information and delay spread [13]. An enhanced XPD model is needed that takes into consideration the elevation departure and arrival angles.

Modifications to Channel Generation

A framework for the generation of the 3D fading channel can be developed based on well-known 2D channel models, such as the IMT-Advanced channel model defined by the ITU Radiocommunications Standardization Sector (ITU-R) M.2135. This section will highlight the required modifications to perform this 3D extension [14].

General Framework

As shown in Fig. 3, the channel generation process for the ITU-R GSCM [14] can be described using six steps. These are classified into three phases:

1. UE parameters
2. Generation of large-scale parameter (LSP) propagation parameter
3. Generation of channel impulse response

The yellow highlighted fields represent the parts that require modification when performing the 3D extension. The user parameter part (step 1) is used to set up simulation parameters, such as the type of environment, the numbers of BSs and UEs, the directions and speeds of the UE, and the propagation condition (LOS/NLOS). In this phase, the users can also supply the antenna at both the BS and UE, and the spacing and orientations of the elements. In this case, and in order to fully utilize the 3D model, the imported antenna patterns should be 3D providing information about the gain and polarization in both the azimuth and elevation planes.

The second part of the ITU-R GSCM creation process is the propagation parameter generation, which consists of path loss (PL), shadow fading (SF) calculation, and the generation of the LSPs and small-scale parameters (SSPs) for the channel. The PL is calculated based on a specified propagation condition provided in in step 2. The generation of the LSPs includes the generation of different channel parameters based on a pre-defined probability distribution function (PDF) with specific mean and standard deviation. These include the root mean square (RMS) delay spread (DS), the RMS AOA and AOD in both azimuth and elevation, the *K*-factor, and the SF. The de-correlation distances and cross-correlations are calculated for the generated LSPs. For more details on these parameters please refer to [14].

The SSPs are now generated based on the LSPs from step 3. The SSPs represent the information associated with each MPC. These include the phase, delay, angular information for each individual cluster, and ray within the cluster. This is performed based on the predefined PDFs. In the proposed 3D extension to the ITU generic channel model, the elevation plane is added in the modeling of the rays' angles. Step 5 represents the generation of the channel impulse response in the time domain. This includes generating random phases for the rays within the cluster, and apply the cross-polarization effect between antenna elements. Then the Doppler effect is added in case of mobility.

Finally, in step 6, path loss and SF values are applied to the channel impulse responses. This stage enables system-level studies to be performed.

Review of the Proposed 3D Channel Models

Having described the principles of 3D channel modeling in the previous sections, we now present a review of our proposed 3D channel model. The authors contributed by modifying the existing 2D ITU generic channel model [14] to include the proposed elevation angle statistics. Channel statistics for many LSPs are generated from a validated ray tracer engine's predictions [12]. Ideal isotropic antenna patterns are applied during the channel predictions stage to decouple the antenna system from the channel model. At a later stage in the system-level study, any BS/UE antenna patterns can be applied as the spatial-phase-polarization convolution process. The proposed channel sta-

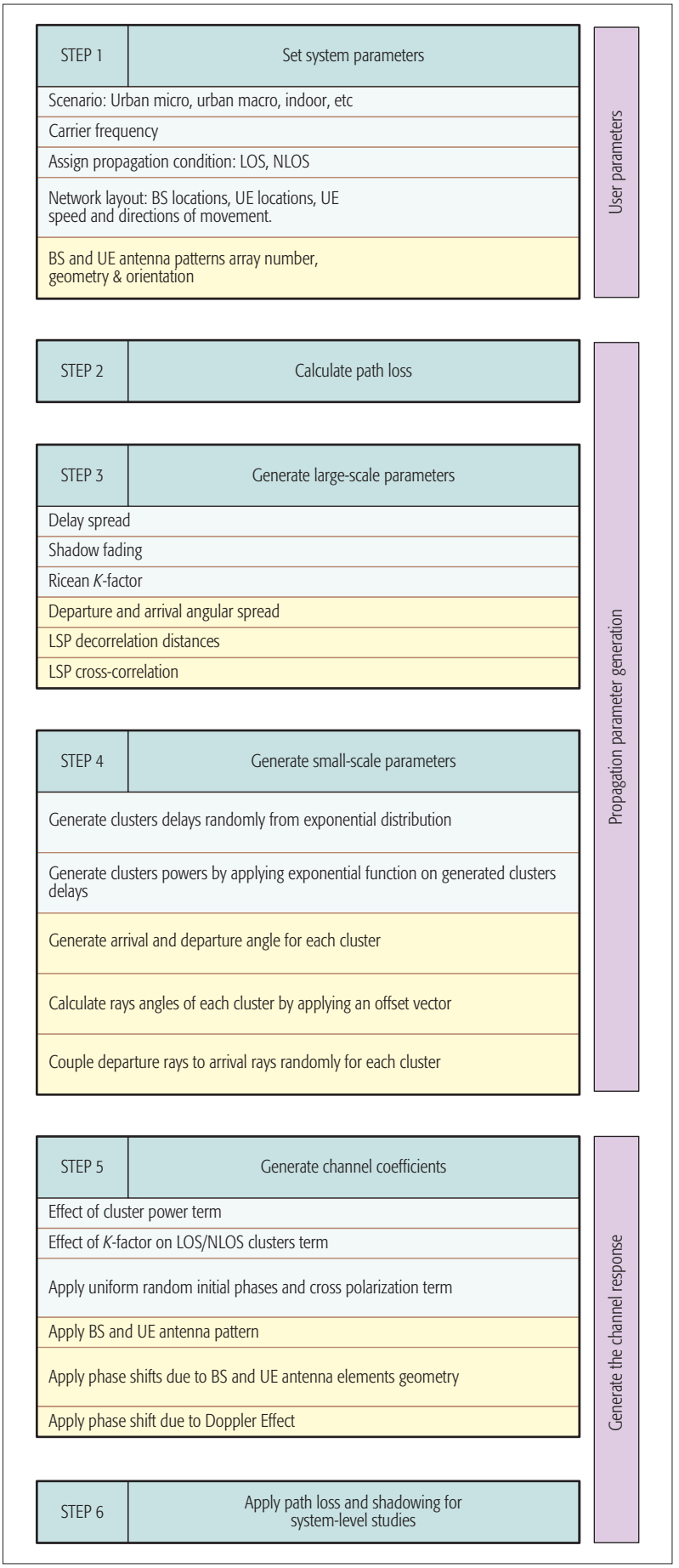


Figure 3. 3D generic channel generation process.

A measurement campaign with a focus on characterising the third dimension of the wireless channel has been conducted in the city centre of Ilmenau, Germany. This was performed in the framework of the WINNER+ project and the measured data determined the resulting model for urban environments.

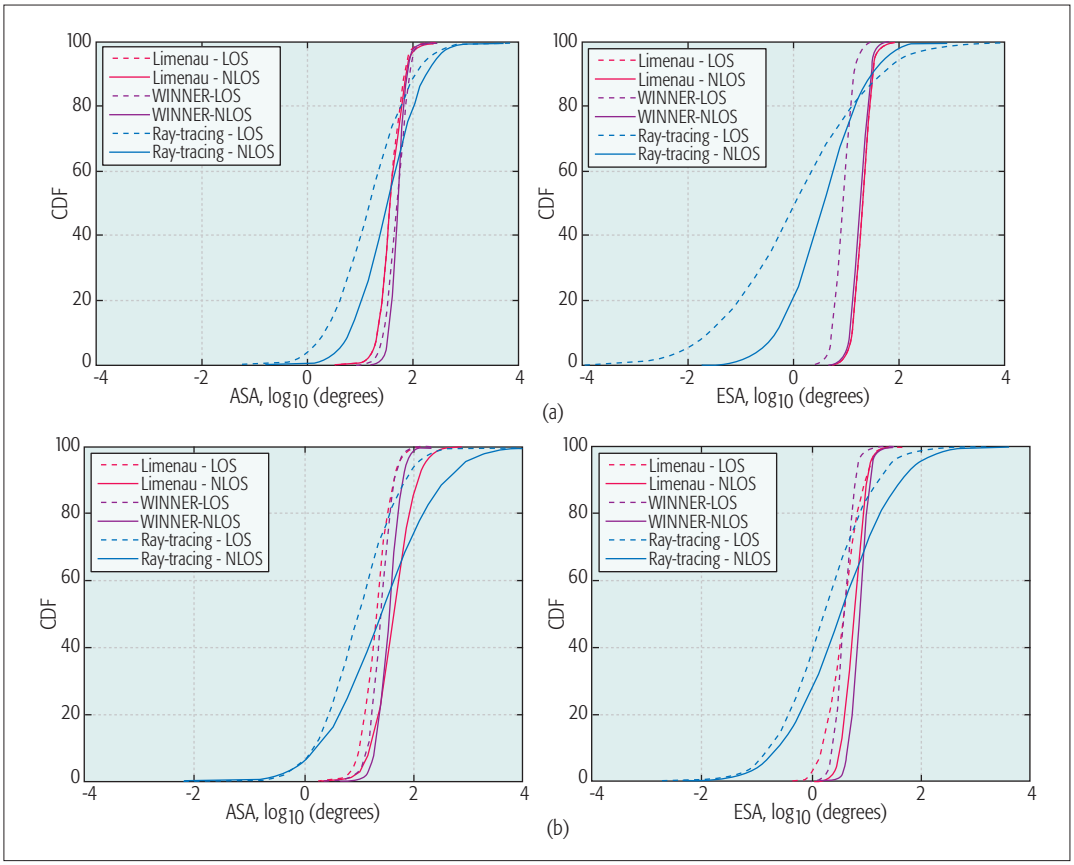


Figure 4. Comparison of CDFs of RMS arrival angular spreads: a) urban macrocell; b) urban picocell.

tistics provide modeling of the PL, SF, and angular spread in both azimuth and elevation planes. The model also covers the de-correlation distances and the cross-correlation for the LSPs. The objective is to calculate the mean (μ) and standard deviation (σ) of the log of all the LSPs based on the distribution functions assumed in the current 2D ITU channel model. An open source code of the enhanced channel model is published in http://enhanced-3d-itu-channel-model.sourceforge.net. Readers are referred to [15] for detailed discussion of the LSP modeling process and the 3D tracer related configurations. The proposed channel models are for macro and microcell environments for different carrier frequencies such as 800 MHz, 2.4 GHz, and 5.9 GHz, and for both LOS and NLOS propagation conditions. It is worth mentioning that the proposed statistics were obtained by averaging all the channel predictions for London and Bristol, United Kingdom.

A measurement campaign with a focus on characterizing the third dimension of the wireless channel has been conducted in the city center of Ilmenau, Germany. This was performed in the framework of the WINNER+ project, and the measured data determined the resulting model for urban environments. Figure 4 presents the cumulative distribution functions (CDFs) of the RMS arrival angular spreads for the WINNER+ urban channel model, the Ilmenau measurements, and our ray-tracing-based channel model. The latter shows the data for macrocells at 2.6 GHz and for lamppost-mounted picocell BSs in Bristol. As expected, the Ilmenau measurements match well to the WINNER+ model, but clear differences can be noticed with the ray-tracing results for Bristol. This highlights a disadvantage of an "one-size-fits-all" empirical channel modeling approach. A possible reason for these differences in the elevation angle statistics are the different city layouts and propagation environments (e.g., the city of Bristol is much more hilly and densely built than Ilmenau), details of the ray-tracing database (e.g., the ray-tracer does not consider parked cars), and the limited number of measured links. In addition, the ray-tracing predictions are based on isotropic antenna patterns; therefore, no filtering of the MPCs is done at the UE side, which leads to higher angular spread compared to the WINNER+ model and the Ilmenau measurements. For the latter, particular antennas were used during the measurements [10], the radiation patterns of which resulted in inevitable spatial filtering of the MPCs. Finally, it should be noted that the distance dependency of the elevation angle spread noted in the ray-tracing data is consistent with the 3GPP observations in [11].

Impact on System-Level Performance

In this section, the impact of the 3D component (elevation) on system-level performance is discussed. Emphasis is placed on comparison with the legacy 2D model.

Doppler Shift

The Doppler frequency component, which is one of the parameters that characterizes the small-scale temporal fading, depends on the AOAs at the UE and the UE velocity vector. The

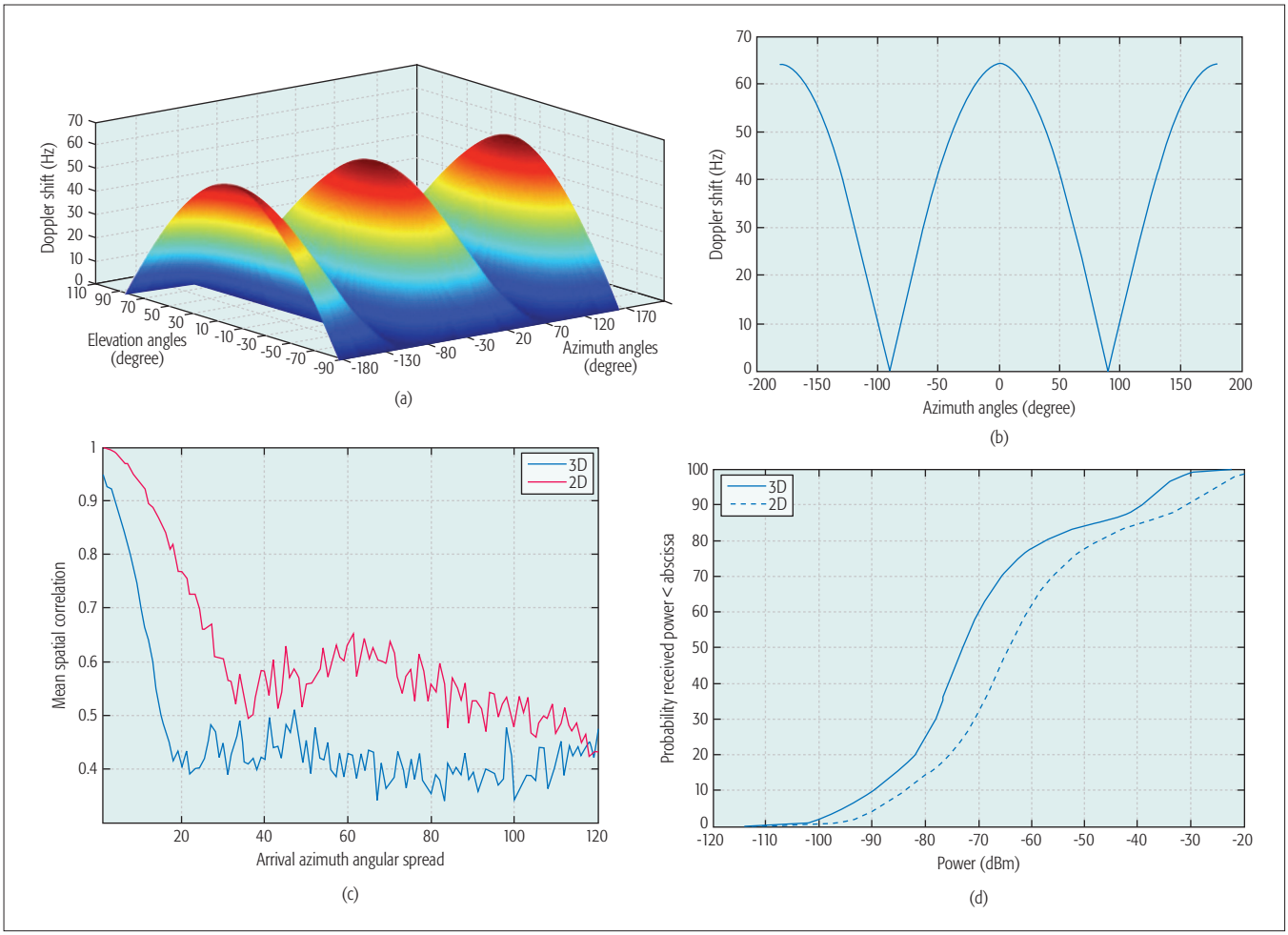


Figure 5. Comparison of 2D and 3D models at system-level performance: a) Doppler shift in 3D model; b) Doppler shift in 2D model; c) comparison of MIMO spatial correlation; d) received power.

calculation of this vector is different in the case of a 3D channel model. For 3D models the elevation angles of the multipath components also affect the Doppler shift (in addition to the azimuth angles already present in 2D models). Based on the ITU model, the Doppler component is calculated as [10]

$$v_{n,m0}(2D) = \frac{||v||\cos(\varphi_{n,m} - \varphi_v)}{\lambda_0} \quad (1)$$

In the 3D model the Doppler frequency is calculated as

$$v_{n,m0}(3D) = \frac{||v||\cos(\varphi_{n,m} - \varphi_v)\cos(\zeta_{n,m} - \zeta_v)}{\lambda_0} \quad (2)$$

where v, λ_0 are the UE velocity (m/s) and wavelength (m), respectively. φ_v is the UE direction of travel in the azimuth plane, and ζ_v is the UE direction of travel in the elevation plane. The variables n, m refer to the sub-path rays in the cluster-based channel model. This difference in the calculation of the Doppler shift affects the spatial fading experienced by UEs moving in a 3D environment. Based on Eqs. 1 and 2, the overall Doppler shift experienced by UEs for different azimuth and elevation angles are presented in Figs. 5a and 5b. These calculations assume absolute UE velocity of 30 km/h.

MIMO Spatial Correlation

When MIMO techniques are deployed, large capacity gains can only be realized when the sub-channels are spatially de-correlated. However, in many real-world propagation environments, the theoretical gains are not achieved due to the significant spatial correlation present in the channel. Our observations show that the 2D model clearly overestimates the level of spatial correlation when the elevation angle is not modeled. Exploiting the elevation plane can further enhance system performance by benefiting from the richness of the 3D channel. The deployment of a 3D channel model requires the application of 3D antenna patterns in the channel generation process. In order to show the difference in spatial correlation between 2D and 3D channel models, Fig. 5c demonstrates the mean correlation at the UE for a range of azimuth angular spreads based on random channel generation for a large number of UEs based on the developed 3D ITU model and the existing 2D ITU model. It is clearly shown that the 3D channel model results in lower spatial correlation.

Total Received Power

Propagation in 3D also impacts the total received power, especially when 3D antenna patterns are included in the analysis. The CDF of the total received power for the 2D and 3D channel mod-

Three-dimensional MIMO (3D MIMO) is an effective step toward massive MIMO, without the need to employ vast numbers of antenna elements at the BS. The vertical dimension can be utilized by the antenna array, with the down tilt of the antenna becoming a significant channel parameter.

els are shown in Fig. 5d. A difference in the total received power between the 2D and 3D models is observed. This data assumes the antenna patterns and the propagation conditions deployed in [15]. The 2D model results in higher power levels compared to the 3D model. This is because the arrival rays in the 2D ITU model are always interpolated at fixed elevation angle. In the 3D ITU model, the rays' elevation dimension is considered, and the interpolation of the antenna gain is considered at all azimuth and elevation angles. At some elevation angles the antenna gain is high, while at other angles the gain is much lower.

3D MIMO and Beamforming

3D MIMO is an effective step toward massive MIMO, without the need to employ vast numbers of antenna elements at the BS. The vertical dimension can be utilized by the antenna array, with the down tilt of the antenna becoming a significant channel parameter. A typical 2D antenna is used to cover a sector of 120° in the horizontal

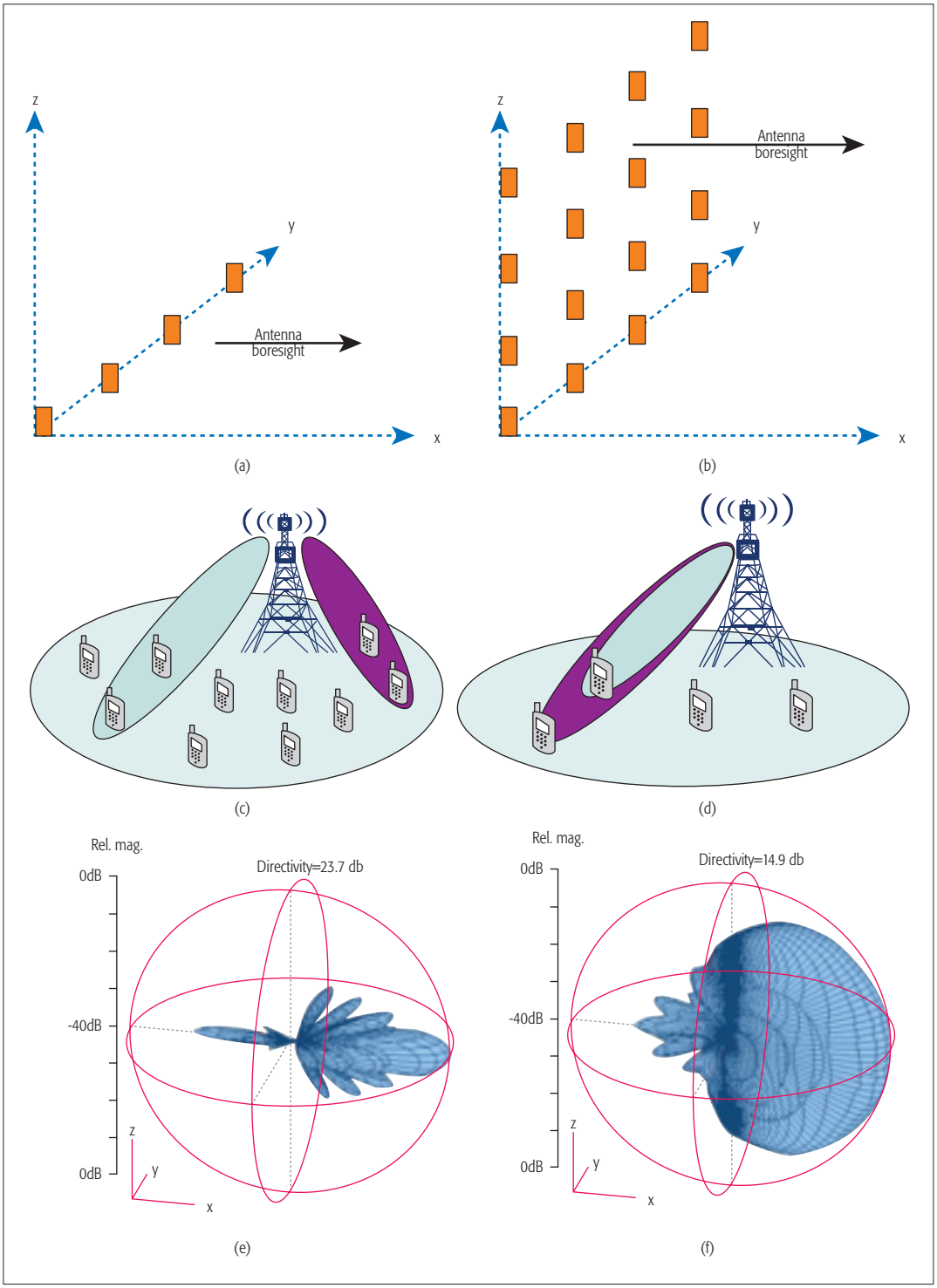


Figure 6. 2D vs 3D antenna beamforming: a) 2D antenna arrangement; b) 3D antenna arrangement; c) 2D beamforming; e) 3D antenna beamforming; f) 2D antenna beamforming.

plane. Compared to 2D channel propagation, in a 3D model the scatterers are no longer located in the same plane as the antenna elements. The incorporation of the elevation plane in the modeling of the departure and arrival angles opens up more opportunities for 3D MIMO, where antenna elements are deployed and spaced in the azimuth and elevation planes. Figures 6a and 6b demonstrate the concept of placing MIMO antenna elements in 2D and 3D space, respectively.

From an antenna perspective, exploiting the elevation plane (as well as the azimuth plane) is commonly referred to as 3D beamforming [2] or FD-MIMO. One way to exploit the additional degree of freedom offered by 3D channels is to adapt the beam pattern for each UE in the vertical direction, thereby improving the signal strength at the receiver while also reducing interference to other UEs. This is unlike the beamforming achieved with a linear array antenna in the horizontal dimension, which does not give full-free space gain, as illustrated in Figs. 6c and 6d. The differences between beamforming in 2D and 3D planes are clarified in Figs. 6e and 6f. These figures show that 3D beamforming results in higher directive gain and narrower beam widths toward a desired point in 3D space (defined by its azimuth and elevation angles).

The higher directive gain obtained from 3D MIMO, when combined with an efficient beamforming algorithm, results in higher system performance compared to azimuth-only beamforming. Analog 3D beamforming represents a promising technology for 5G systems for coverage enhancements and inter-cell interference cancellation.

Conclusions

This article has presented an overview of some the research in 3D channel modeling and has highlighted the relevant contribution of some standardized channel models. It has also discussed the benefits of the 3D model and its impact on system-level performance. The proposed 3D channel model is presented with explanation of the modeled channel parameters. To illustrate the subjects covered in this article the authors have released open source code at sourceforge.net that implements the proposed 3D extension to the 3GPP/ITU model.

References

[1] J. Fuhl, J.Rossi, and E. Bonek, "High-Resolution 3-D Direction-of-Arrival Determination for Urban Mobile Radio," *IEEE Trans. Antenna and Propagation*, vol. 45, 1997, pp. 672–82.

[2] Z. Hu *et al*., "Work in Progress: 3D Beamforming Methods with User-Specific Elevation Beamfoming," *9th Int'l. Conf. Commun. and Networking in China*, 2014, Aug. 2014, pp. 383–86.

[3] T. Aulin, "A Modified Model for the Fading Signal at a Mobile Radio Channel," *IEEE Trans. Vehic. Tech.*, vol. 28, 1979, pp. 182–204.

[4] A. Juchar, J. Rossi, and E. Bonek, "Directional Macro-Cell Channel Characterization from Urban Measurements," *IEEE Trans. Antenna and Propagation*, vol. 48, 2000, pp. 137–46.

[5] K. Kalliola, K. Sulonen, and H. Laitinen, "Angular Power Distribution and Mean Effective Gain of Mobile Antenna in Different Propagation Environments," *IEEE Trans. Vehic. Tech.*, vol. 51, 2002, pp. 823–38.

[6] J. Laurila, K. Kalliola, and M. Toelstch, "Wide-band 3-D Characterization of Mobile Radio Channels in Urban Environment," *IEEE Trans. Antennas and Propagation*, vol. 50, 2002.

[7] K. Kalliola *et al*, "3-D Double-Directional Radio Channel Characterization for Urban Macrocellular Applications " *IEEE Trans. Antenna and Propagation*, vol. 51, 2003, pp. 3122–33.

[8] L. Hentila, "Elevation Extension for a Geometry-Based Radio Channel Model and Its Influence on MIMO Antenna Correlation and Gain Imbalance," presented at the 5th Euro. Conf. Antenna and Propagation, 2011.

[9] R1-122034, "Study on 3D Channel Model for Elevation Beamforming and FD-MIMO Studies for LTE," 3GPP TSG RAN2012.

[10] WINNER+ Final Channel Models, D5.3 V1.0, June 2010.

[11] 3GPP TR 130500, "Detailed 3D Channel Model for LTE," 3GPP, St.Julian's Malta 3GPP TSG RAN Meeting #72, 2013.

[12] E. K. Tameh and A. R. Nix, "The Use of Measurement Data to Analyse the Performance of Rooftop Diffraction and Foliage Loss Algorithms in A 3-D Integrated Urban/Rural Propagation Model," *IEEE VTC-Spring*, May 1998.

[13] Q. Nadeem, A. Kammoun, and M. Alouini, "A Generalized Spatial Correlation Model for 3D MIMO Channels Based on the Fourier Coefficients of Power Spectrum," *IEEE Trans. Signal Processing*, Apr. 2015.

[14] ITU-R M.2135-1, "Guidelines for Evaluation of Radio Interface Technologies for IMT-Advanced," Dec. 2009.

[15] R. Almesaeed *et al*, "A Proposed 3D Extension to the 3GPP/ITU Channel Model for 800MHz and 2.6GHz Bands," *8th IEEE EUCAP*, Apr. 2014.

Biographies

Reham Almesaeed received her Ph.D and M.S.c from the University of Bristol, United Kingdom, in 2017 and 2010, respectively. She is currently a lecturer in the Computer Engineering Department at the University of Bahrain. Her research interests include wireless channel modeling, wireless local area networks, Long Term Evolution, fifth-generation communication systems, mmWave communications, and massive MIMO.

Araz Sabir Ameen received his Ph.D. degree in electrical and electronic engineering from the University of Bristol in 2015. Currently he is a lecturer in wireless communication at the University of Sulaimani, Kurdistan, Iraq. His research interests include orthogonal frequency-division multiplexing, vehicular communications, multiple antenna systems, Long Term Evolution-Advanced and fifth-generation communication systems, adaptive 3D beamforming, wireless channel modeling, intercell interference modeling, and mitigation techniques.

Evangelos Mellios is currently a lecturer in electrical and electronic engineering at the University of Bristol. He joined the Communication Systems and Networks Research Group where he completed his Ph.D. in 2013, and he subsequently worked as a research assistant in on-body sensing and wireless communications within the SPHERE Project (a Sensor Platform for Healthcare in a Residential Environment).

Angela Doufexi received her Ph.D. degree from the University of Bristol in 2002. She is currently a reader in wireless networks at the University of Bristol. Her research interests include orthogonal frequency-division multiplexing, multiuser diversity and resource allocation, wireless local area networks, vehicular communications, multiple antenna systems, Long Term Evolution and fifth-generation communications systems, mmWave communications, massive MIMO, and multimedia transmission. She is the author of over 150 journal and conference papers in these areas.

Andrew Nix received his B.Eng and Ph.D. degrees from the University of Bristol in 1989 and 1992, respectively. He is currently a professor of wireless communication systems and Dean of Engineering at the University of Bristol, where he leads the Communication Systems and Networks research group. His research interests include 5G networks and architectures, connected and autonomous vehicles, radio wave propagation modeling, adaptive antenna arrays, massive MIMO, and advanced Wi-Fi systems. He has supervised more than 50 Ph.D. students and published in excess of 400 journal and conference papers.

The higher directive gain obtained from 3D MIMO, when combined with an efficient beamforming algorithm, results in higher system performance compared to azimuth-only beamforming. Analogue 3D beamforming represents a promising technology for 5G systems for coverage enhancements and inter-cell interference cancellation.

ACCEPTED FROM OPEN CALL

Blended Antenna Wearables for an Unconstrained Mobile Experience

Matilde Sánchez-Fernández, Antonia Tulino, Eva Rajo-Iglesias, Jaime Llorca, Ana García Armada

As a catalyzer for the capacity increase in both directions of the wireless communication link, we propose an innovative idea that brings the spectral and energy efficiency benefits of massive MIMO systems directly to the end user without compromising device size, weight, or power consumption. We propose a radio enhanced garment composed of blended textile antennas for seamless high data rate connectivity anytime, anywhere, addressing immediate and future needs.

ABSTRACT

We envision a world in which everyday life experiences can be augmented on-demand via the real-time cloud processing of information sourced at multiple wireless end points. While current wireless systems focus their effort on improving downlink capacity, these life-changing augmented experiences will only become a reality if the uplink capacity increases at the same or even higher rate. As a catalyzer for the capacity increase in both directions of the wireless communication link, we propose an innovative idea that brings the spectral and energy efficiency benefits of massive MIMO systems directly to the end user without compromising device size, weight, or power consumption. We propose a radio enhanced garment composed of blended textile antennas for seamless high data rate connectivity anytime, anywhere, addressing immediate and future needs. The real-world applications for such a solution are tremendous, including enhanced connectivity in crowded spaces and remote areas, as well as symmetric extremely high data rates for access to next generation real-time services (e.g., augmented reality and cognition, real-time computer vision, telepresence, 3D video sharing) from ever lighter end user devices (e.g., Google Glass).

INTRODUCTION

In a fast approaching future, Internet traffic will be dominated by the consumption of resource intensive and interaction intensive applications (e.g., augmented reality, real-time computer vision, immersive and interactive 3D video) running in distributed cloud nodes and accessed from a massive number of resource limited wireless user devices (e.g., smart phones/tablets/watches/glasses/wearables). The efficient and sustainable evolution toward this attractive future will call for disruptive innovations that ensure extremely high uplink and downlink data rates without affecting the desirable small size, lightness, and seamless properties of end user devices.

Next generation wireless communication system requirements are driven by these new bandwidth-intensive real-time applications. Fifth generation (5G) cellular networks envision 100-fold capacity gains, simultaneous connections for billions of devices, and a 10 Gb/s individual user experience with extremely low latency and response times. It is obvious that no single technology will address all these challenges and that this radical change of performance will significantly modify the communication system at all levels [1].

In this work, with the aspiration of becoming one of the key enablers of the envisioned 5G 100-fold symmetric capacity growth, we propose a technology that directly addresses the needs of the most challenging and critical component of the communication system, the user end point. Any solution that aims at boosting the capacities of user devices' cannot compromise their seamless, lightweight, portable features. The deployment of a large number of antennas at the user end would be a candidate technological solution for improving system capacity if we could go beyond the desirable reduced dimensions of the end device. Our approach exploits the user device's closest surroundings, i.e., the user's own clothing or accessories (e.g., laptop mats, bags, suitcases), to increase the device's capabilities without compromising its lightweight, portability, and energy efficiency features. Specifically, we propose to deploy a large number of antennas at the user end, by blending textile antennas around the user's personal sphere (e.g., clothing, accessories). Such an antenna-based wearable would connect to any data-enabled user device, immediately boosting its communication capabilities. This technological solution, whose simulated performance, design alternatives, and feasibility are explored in detail in this work, shows very promising results in terms of achievable data rates. Figure 1 shows that 40 textile antennas attached to a user device for uplink transmission can provide high data rate symmetric connectivity, enabling future wireless systems aligned with LTE-evolution demands. Observe that with today's available uplink bandwidth, a single-antenna terminal can only achieve data rates on the order of tens of Mb/s, enabling applications that would not go beyond current IP television (IPTV). Increasing the number of antennas at the user end enables services such as gaming, cloud computing, or even 3D high definition (HD) video, whose bandwidth needs move to the several hundreds of Mb/s. This article introduces MIMOmat, a disruptive blended antenna wearable technology with the potential to dramatically boost the mobile user experience. We provide a comprehensive overview of this breakthrough technology that starts with a review of existing high capacity-achieving technologies, and then cover blended multi-antenna design,

This work has been partly funded by the Spanish Government through projects MIMOTEX (TEC2014-61776-EXP), CIES (RTC-2015-4213-7), ELISA (TEC2014-59255-C3-3R), and TEC2013-44019-R.

Digital Object Identifier: 10.1109/MCOM.2017.1500722

Matilde Sánchez-Fernández, Eva Rajo-Iglesias and Ana García Armada are with Universidad Carlos III de Madrid; Antonia Tulino is with the University of Napoli and Nokia-Bell Labs. Jaime Llorca is with Nokia-Bell Labs.

integration within the cellular infrastructure, spectral efficiency performance quantification, and future challenges.

High Capacity-Achieving Technologies

There are three fundamental dimensions that can be explored for increasing the capacity of wireless communication systems:

- Space (e.g., via network densification)
- Spectrum
- Spectral efficiency

A key enabling technology for high spectral efficiency is the use of multi-antenna systems, also known as MIMO [4]. This technology provides a radical increase in capacity that is proportional to the number of radiating elements. However, it is important to note that these gains are constrained by the smaller number of antennas deployed at any extreme of the communication system, up to the point that a system with two antennas at the transmitter and two antennas at the receiver outperforms a system with an extremely large number of antennas at the transmitter and just one antenna at the receiver [5]. The spectral efficiency gains that can be achieved via the deployment of a large number of antennas is hence limited by:

- The inherent increase in signal processing complexity associated with MIMO systems [6]
- The spatial restrictions for the deployment of a large number of antennas at both ends of the communication link.

In a wireless cellular system, the base station (BS) can leverage a number of favorable attributes such as grid-power, signal processing capabilities, and physical space availability. However, the user end has become an obvious bottleneck for potential performance improvement: space restrictions usually apply, available power is highly limited, and non-costly hardware implementations are crucial. Recent studies have proposed the use of a massive number of antennas, also referred to as massive MIMO [6], at the BS. Massive MIMO at the BS can provide multiplicative total throughput gains by matching the number of users served to the number of antennas at the BS. However, the reduced number of antennas at the user end limits the per-user symmetric throughput required to enable next generation real-time applications. Indeed, in spite of massive MIMO advances at the BS, the tight restrictions at the user end become a major obstacle in LTE evolution. MIMOmat is a wearable extension to the user device with blended antennas in textile technology that aims at overcoming this hurdle by also bringing massive MIMO to the user terminal. This solution emerges as a disruptive joint venture between massive MIMO and textile antenna technologies that could provide on-demand symmetric data rate increases unimaginable today.

The deployment of a large number of antennas in the least suitable component of the communication system is a challenging idea that leapfrogs current research in massive MIMO systems. To accomplish an end-to-end solution, not only the design and deployment of the antennas should be addressed, but also the signal processing needs, the RF chains design, and all hardware related issues that might jeopardize the wearability of the solution. We review these key aspects in the following sections.

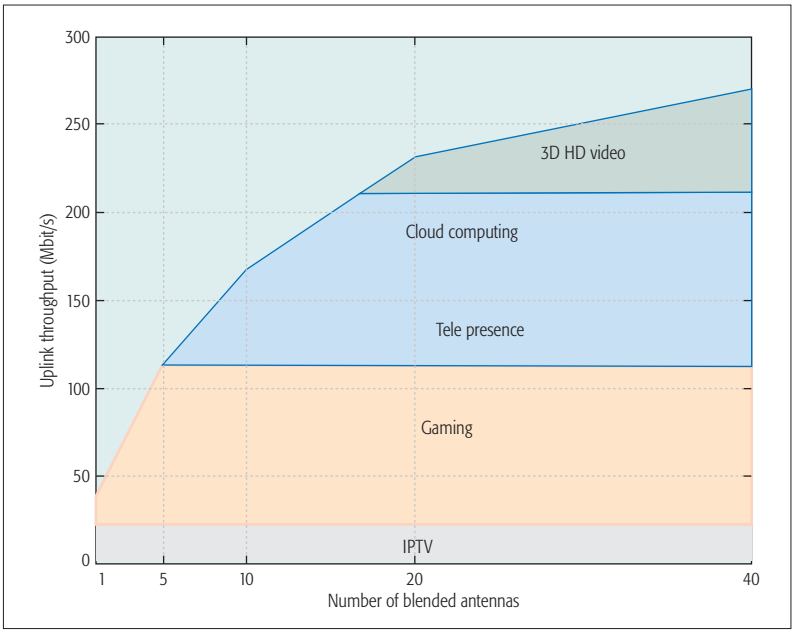


Figure 1. Average uplink throughput obtained via blended textile antennas at the user end for a bandwidth of 5MHz and 64 antennas at the base station in a realistic cellular scenario [2]. The different applications that can be supported are highlighted based on [3].

Seamless Antenna Systems

Textile antenna technology is sufficiently mature to be used in an application where the antennas need to be seamlessly embedded in the user's personal sphere. Indeed, recent studies have proposed the use of single textile antennas for off-body communications [7] and as a MIMO catalyst [8, 9]. However, to the best of our knowledge, textile antenna technology has not been proposed with a large deployment (tens of antennas) view, where open challenges related to radiated power, bandwidth, and the mutual coupling (MC) apply.

MIMOmat is the first technological solution that deploys a large antenna array, miniaturized in wearable form at the user end via embedded textile technology. Our design is intended to be a plug-&-play solution to which any data-enabled device can connect, providing on-demand high data rate, reliable, low power communications (Fig. 2).

Blended Antenna Wearable Design

We propose a tailored textile planar antenna array design as the enabling solution to provide the user terminal with a wearable extension comprising a large number of antennas. As described in [7], for an antenna to be wearable, it is necessary to combine the right selection of materials for both the conductive and the non-conductive components. With this in mind, we use a textile tissue (commonly felt) as a non-conductive substrate, while the metallization is implemented with electrotextile materials (e.g., Shieldit® conductive materials).

The planar antenna array solution was first simulated to design and fine tune the antenna parameters of interest, and later built to ensure its wearability, with special emphasis on weight and flexibility. The simulations were performed using

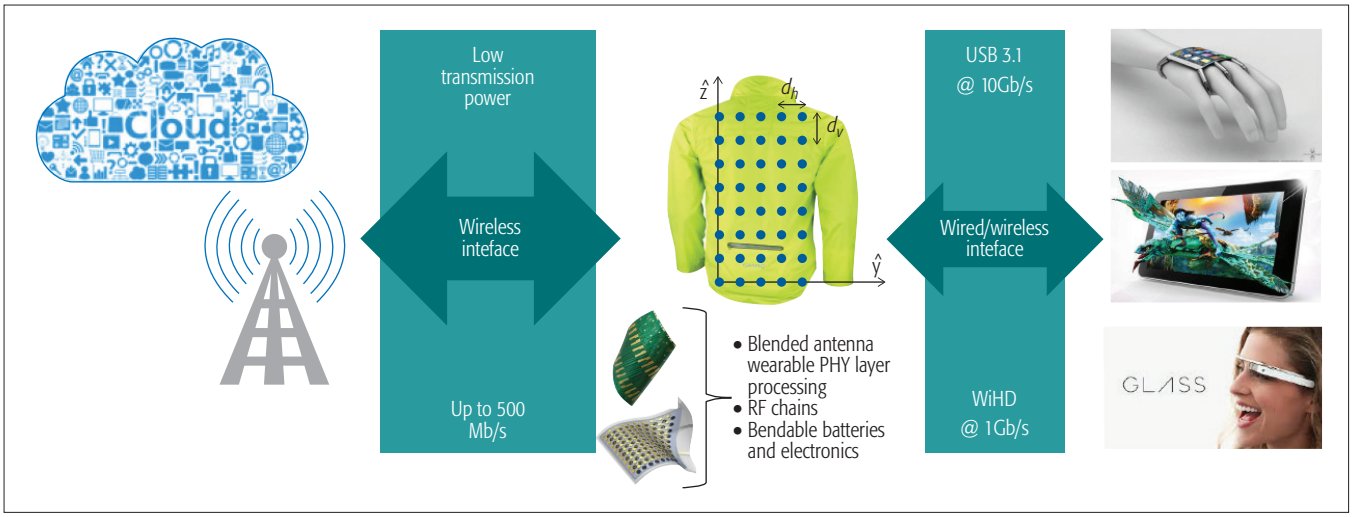


Figure 2. System model. Example of a blended antenna wearable where the antennas are deployed in the back of a jacket.

CST Microwave Studio, a full wave simulation tool that provides radiation patterns (RP), antenna gains, and *S*-parameters (*S* stands for scattering, and such parameters are associated with return loss and voltage standing wave ratio). The *S*-parameters give crucial information about the antenna system performance: the antenna matching S_{ii} dictates the individual matching frequency and the bandwidth of the *i*-th antenna, while the S_{ij} parameter determines the MC between the *i*-th and the *j*-th antenna.

We designed squared patch antennas of length 46.5 mm using a 3 mm thick felt (with dielectric constant $\in_r = 1.38$) as substrate. The simulations (Fig. 3a) indicate that each individual antenna is well matched to a transmission center frequency of 2.5 GHz and exhibits a bandwidth of 70 MHz (defined as the range of frequencies for which the reflection coefficient S_{ii} is below or equal to –10 dB).

We remark that, while the proposed solution is designed with a central frequency of 2.5 GHz and a bandwidth of 70 MHz, both the antenna central frequency and the antenna bandwidth can be easily tuned for the specific application. It should be noted that different frequency bands are being explored for 5G, among which, the band below 6 GHz is intended to be used for LTE co-existence. In the proposed central frequency, the antenna design would approximately cover LTE band #7 if operating in FDD mode, or band #41 if operating in TDD mode [10].

For the planar array geometry, we chose an 80 mm inter-element distance (0.66λ) to minimize the MC among antennas, and deployed the squared patch antennas following $M_h = 5$ columns in the horizontal plane and $M_v = 8$ antenna rows in the vertical plane (Fig. 3b). The total planar array size is $M = M_v \times M_h = 40$ antennas occupying 44.6×60.6 cm^2. Such a planar antenna array could be deployed, for example, in the back of a jacket, as shown in Fig. 2, or in a flexible and lightweight textile mat that could easily fit in a bag. When jointly deploying the antennas in a planar array, the RPs of the individual antennas slightly vary (see the different RP shapes in Fig. 3b) with antenna gains going from 7.58 to 8.05 dB depending on the relative position of the antennas. It is important to note that the radiation to the body is minimal, as clearly shown in Fig. 3b with negligible radiation below the planar array. In addition, the simulated MC among antennas, determined by the S_{ij} parameters, is always below –20 dB (Fig. 3a).

Finally, the antennas were built with the same materials and geometry chosen in the CST simulation stage. The actual deployment confirmed the lightweight properties of the solution, with 3 grams per antenna, and the high flexibility provided by the felt and the electrotextile material.

Blended Antenna Wearables in Next Generation Cellular Networks

The integration of MIMOmat into a cellular infrastructure can be devised in Fig. 2. Here, the blended antenna wearable relies on two important interfaces: one interface between the user device and the blended antenna wearable (a jacket in Fig. 2), and a second interface between the wearable and the cellular network.

The first interface connects a data enabled device (e.g., phone/tablet/watch) with the massive MIMO blended antenna wearable. The device-wearable interface should support the high data rates envisioned to be delivered via the MIMOmat. A feasible wired interface, in terms of connection rates, would be based in USB 3.1 with a transmission rate of 10 Gb/s. The wireless HD technology interface (WiHD), designed to enable wireless streaming of high-definition multimedia between source devices and displays, could be a second option. The link rate supported by this technology is 1 Gb/s. Both technologies have their pros and cons and the interface choice should also take into account the complexity burden (in terms of specific hardware and signal processing requirements) on the wearable implementation. The second interface connects the massive antenna wearable to the wireless network, and it is the foundation for the capacity increase experienced by the end user. This interface needs to define a novel physical layer (PHY) that will eventually override the original physical layer of the user device. This PHY layer should specify duplexing schemes for uplink-downlink communication, synchronization, channel esti-

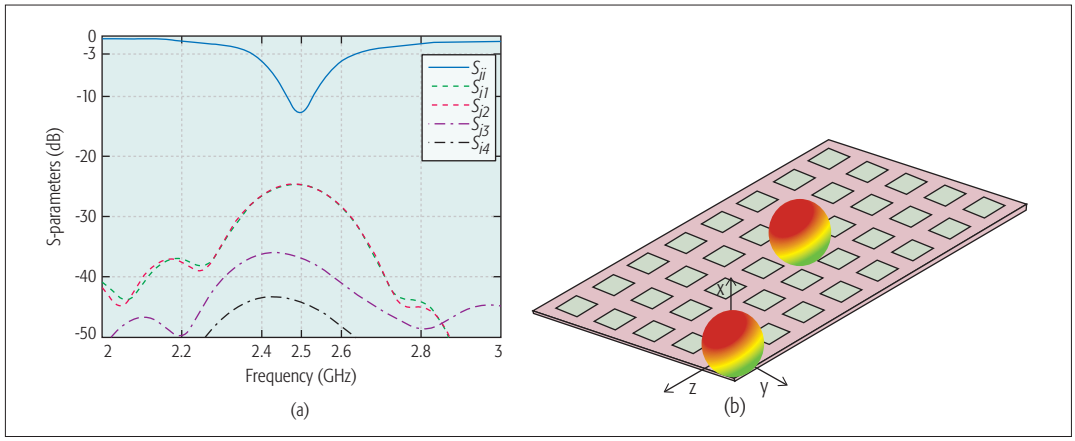


Figure 3. Textile antenna planar array simulated performance: a) *S*-parameter magnitude values for antenna placed approximately in the center of the planar array. The antenna matching is given S_{ii} and also the mutual coupling with the four antennas surrounding up, down, left, right (S_{ij}); b) deployed antenna set and example of radiation pattern.

The MIMOmat, with existing wireless/wired interfaces to connect to the user device, and novel PHY protocols to connect to the wireless network, becomes a plug-&-play solution to which any data-enabled device can connect in order to enhance its connectivity and data throughput.

mation, signal processing, and all the required signaling to be compliant with the BS PHY. The associated electronics and powering components included in the wearable should feature lightweight, flexibility, and low power consumption requirements [11]. The rest of the communication protocols (over PHY) shall be supported by the device that is connected to the wearable extension.

In summary, the MIMOmat, with existing wireless/wired interfaces to connect to the user device, and novel PHY protocols to connect to the wireless network, becomes a plug-&-play solution to which any data-enabled device can connect in order to enhance its connectivity and data throughput.

In the following section, we assess the potential impact of MIMOmat on users' mobile experience by identifying the tradeoff between the complexity burden of PHY functionality on the wearability of the MIMOmat and the achievable data rates.

Unconstrained Mobile Experience

Today's wireless data traffic is dominated by multimedia content delivery services such as on-demand audio and video streaming, which require high data rates in the downlink. The uplink capacity is limited, not only because of the technological constraints (e.g., size, power, computational complexity) inherent in mobile equipment already mentioned in previous sections, but also because most current and historical data communication services have not demanded high bandwidth requirements in this direction. Asymmetric capacity assumptions between downlink and uplink have therefore largely characterized the design and implementation of wireless communication systems. Blended antenna wearables are envisioned to break this tendency and become one of the key 5G enablers (Fig. 1) providing reliable symmetric high data rates to/from any user device.

Think of a user trying to enjoy an augmented reality application from a lightweight video-enabled device. The reduced uplink data rate limits the service to 2D video with a few augmented tags that identify a subset of objects in the captured scene. As soon as the user connects the device to their MIMO-enabled wearable, the application delivers 3D video with real-time full scene analysis, a dramatic and instantaneous user experience boost.

In this section, in order to quantify the disruptive performance improvements that can be achieved via MIMOmat, we focus on two simulated scenarios:

- An idealistic scenario with a single user equipped with a MIMOmat transmitting to a BS
- A realistic cellular network scenario with the presence of multiple interfering users

For both scenarios, we evaluate the average achievable uplink rates. To this end, we first describe the channel model and the possible transmission strategies, and then quantify the achievable data rates for each scenario.

Channel Model and Parametrization

The narrowband massive MIMO channel is described by a $N \times M$ channel matrix $\mathbf{H}$. The channel matrix is characterized by the antenna array geometry, radiation patterns in transmission and reception, and the surrounding scattering environment. The element h_{nm} in the channel matrix $\mathbf{H}$ is described by Green's function [12] sampled at the position of the n-th receiving antenna $\mathbf{r}'_n$ given that the point source is located at the m-th transmitting antenna ($\mathbf{r}_m$):

$$h_{nm} = \iint \mathcal{G}_m(\theta,\phi)\mathcal{G}'_n(\theta',\phi')S(\mathbf{k}'(\theta',\phi'),\mathbf{k}(\theta,\phi)) \\ e^{-j\mathbf{k}(\theta,\phi)\mathbf{r}_m} e^{j\mathbf{k}'(\theta',\phi')\mathbf{r}'_n} d\mathbf{k}'(\theta',\phi') d\mathbf{k}(\theta,\phi) \quad (1)$$

In Eq. 1, $\mathcal{G}_m(\theta, \phi)$ and $\mathcal{G}'_n(\theta', \phi')$ are the radiation patterns in azimuth (ϕ) and elevation (θ) at the transmitter and receiver, respectively; $\mathbf{k}(\theta, \phi i)$ and $\mathbf{k}'(\theta', \phi')$ represent the wave vector space at the transmitter and receiver, respectively; and $S(\mathbf{k}'(\theta', \phi'), \mathbf{k}(\theta, \phi))$ is the channel scattering function, which relates the plane wave's emitting and receiving directions, $\mathbf{k}$ and $\mathbf{k}'$, respectively. It should be noted that there is no dependence on time of the position of the antennas, $\mathbf{r}'_n$ and $\mathbf{r}_m$, and hence we are considering a static scenario with no block or relative movement of the antennas.

In order to quantify the achievable rates for the two described scenarios, the channel model in

It should be noted that we are not only taking into account the positive increase in the matrix dimension that is generated by the fact that we add more antennas at the user end, but we are also considering the antennas' non-ideal characteristics or impairments such as non-broadside RP, mutual coupling, and antenna gains.

Eq. 1 needs to be parameterized with the antenna design parameters obtained earlier. Accordingly, the simulated RPs of the wearable antennas are used for $\mathcal{G}_m(\theta, \phi)$, and the proposed planar array geometry with 0.66λ separation is used for the antenna positions $\mathbf{r}_m$. Recall that the proposed design contains $M = 40$ antennas. However, in order to highlight the benefits of this massive $M = 40$ antenna deployment, we also consider planar arrays with fewer antennas, $M = \{1, 5, 10, 20, 40\}$, all with inter-element distance of 0.66λ and with their corresponding measured RPs. In terms of mutual coupling, while all measured MC values were bellow −20 dB and could be easily neglected, we still include them in our simulations. For the sake of simplicity, we consider ideal antennas (broadside RP) with zero mutual coupling at the BS (receiving end), i.e., $\mathcal{G}'_n(\theta', \phi') = 1\ \forall\phi, \theta$ values. The BS deploys $N = \{1, 4, 8, 64\}$ antennas in a linear array with a distance among antennas of λ/2, except for $N = 64$, where we assume an 8×8 planar array with the same inter-element separation, providing the receiving antenna positions $\mathbf{r}'_n$.

For the scattering function parameterization $S(\mathbf{k}'(\theta', \phi'), \mathbf{k}(\theta, \phi))$, we follow standard approaches in the literature. At the user end, we assume full angular dispersion that we shall model uniformly. At the BS, the angular spread could be significantly smaller than at the user end, depending on the position of the base station. In order to get a worst case performance, we assume a narrow angular spread in azimuth (≈ 30 degrees) and negligible angular spread in elevation.

With all these parameters, channel matrix samples can be generated and used for rate computations. It should be noted that we are not only taking into account the positive increase in the matrix dimension that is generated by the fact that we add more antennas at the user end, but we are also considering the antennas' non-ideal characteristics or impairments such as non-broadside RP, mutual coupling, and antenna gains.

Transmission Strategies in the Wearable

As stated earlier, the signal processing functionality at the MIMOmat is a key component driving the fundamental complexity-performance tradeoff. In the case of a massive antenna wearable designed to boost symmetric data rates, the signal processing complexity is dominated by:

- The precoder computational complexity (e.g., matrix operations, finite precision operations)
- The amount of channel state information (CSI) available at the MIMOmat.

We propose four different precoders, each of them characterized by different computational complexity and available CSI at the transmitter:

- *Optimal precoder under instantaneous CSI (Optimal I-CSI)*: the precoder instantaneously diagonalizes the channel matrix, and its squared singular values are given by the optimal water-filling (WF) power allocation [2].
- *Matched filter (MF)*: the precoder consists of the transpose conjugate of the channel matrix [2].
- *Optimal precoder under statistical CSI (Optimal S-CSI)*: the precoder's eigenvectors diagonalize the channel matrix in an average manner, while its squared singular values are equal to the fraction of average signal power recovered by a minimum mean squared error (MMSE) receiver from the corresponding eigenvectors [13].
- *Optimal precoder with no CSI (No CSI):* the precoder is isotropic, i.e., equal to the identity matrix [2].

Note that the first two precoders require the transmitter to accurately track the instantaneous CSI, which may be feasible with the system working in time division duplexing (TDD) mode, while Optimal S-CSI only requires access to the channel distribution, with less sensitivity to the channel coherence time, and No-CSI does not require any knowledge. In terms of computational complexity, *Optimal I-CSI* requires a matrix decomposition computation at any channel use, while *Optimal S-CSI* only performs such computation at any change of the channel statistics. Note that MF and No-CSI directly use the transport conjugate of the channel matrix and the identity matrix, respectively, and hence do not require any extra computation.

Uplink Achievable Rates in an Idealistic Scenario

In this section, we assess the throughput performance of an idealistic scenario where a single wireless device equipped with a MIMOmat transmits to a BS over a channel described earlier. Specifically, we focus on quantifying the average achievable uplink rate (in bits/s/Hz), which for the Gaussian input signal[1] is given by

$$R = \mathbb{E}\left[\log\det\left(\mathbf{I}_N + \frac{\mathrm{SNR}}{\mathrm{Tr}\{\mathbf{Q}\}}\mathbf{H}\mathbf{Q}\mathbf{H}^{\dagger}\right)\right], \qquad (2)$$

where SNR denotes the transmitting signal-to-noise ratio, $\mathbf{I}_N$ is the $N \times N$ identity matrix, $\mathbf{H}$ is the narrowband MIMO $N \times M$ channel matrix, as described earlier, $\mathbf{Q}$ denotes the input covariance matrix chosen depending on the implemented precoding strategy, and the expectation is taken over the channel distribution. Finally, note that when the achievable rates are simulated for a specific signaling bandwidth W, as in some of the scenarios studied here, the throughput is then given by $W \times R$, and measured in bits/s.

In the following, we plot the average achievable rates as given by Eq. 2 for the four transmission strategies detailed earlier, each corresponding to a different $\mathbf{Q}$. Under the assumption of instantaneous CSI at the transmitter, the input covariance matrix takes the form of $\mathbf{Q} = \mathbf{U}\Lambda\mathbf{U}^{\dagger}$, where $\mathbf{U}$ are the eigenvectors of $\mathbf{H}^{\dagger}\mathbf{H}$, while Λ is the result of the classical WF power allocation algorithm. Under MF precoding, the input covariance matrix takes the form $\mathbf{Q} = \mathbf{H}^{\dagger}\mathbf{H}$. In the statistical CSI scenario, the eigenvectors of $\mathbf{Q}$ are given by the eigenvectors of the channel covariance matrix $\mathbb{E}[\mathbf{H}^{\dagger}\mathbf{H}]$, while its eigenvalues are obtained via the optimal power allocation algorithm described in [13]. Finally, if no CSI is available at the transmitter, $\mathbf{Q} = 1/M\mathbf{I}$.[2]

In this context, we analyze two different scenarios with a varying number of antennas at both ends of the communication system.

In the first scenario, motivated by the limited number of antennas in today's BS, we consider $N = 4$ antennas at the BS and $M = \{1, 10, 20, 40\}$ blended antennas at the user. Note that in this

[1] We remark that under more practical signaling schemes (non-Gaussian signaling), there will only be a loss of approximately 3 dB.

[2] Note that all the transmission strategies are capacity achieving, except the MF precoding. This is why we use the term achievable rate, instead of capacity, throughout the document.

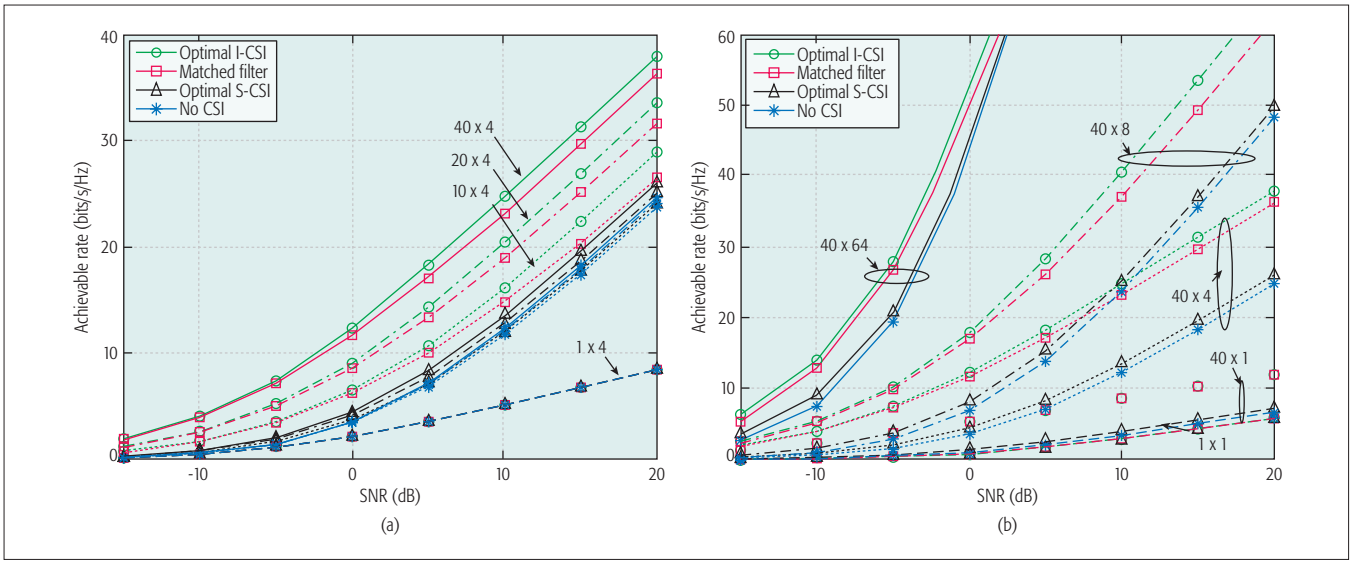


Figure 4. Average achievable rates in the uplink in an idealistic scenario: a) performance trade-off for different number of blended antennas deployed at the wearable $M = \{1,10, 20, 40\}$. The BS antenna configuration is fixed to $N = 4$ representing a nowadays scenario; b) performance trade-off for $M = 40$ blended antennas deployed at the wearable. The BS allows different antenna configuration $N = \{1, 4, 8, 64\}$ representing the evolution of BS in future wireless systems.

case, the potential linear increase of the transmission rate with the number of antennas at the wearable will be limited by the four antennas at the BS. Figure 4a shows that the deployment of an increasing number of blended antennas at the user end has clear benefits in terms of achievable data rates. In this case, the plots show a linear increase of spectral efficiency with logSNR. For a given precoding strategy, the gain can be seen in terms of the spectral efficiency for a fixed SNR, or in terms of SNR for a target achievable rate. Compared to just one antenna, the deployment of 40 blended antennas at SNR = 5 dB gives an increase of up to 4× in achievable data rate.

In the second scenario, in order to illustrate the multiplicative gains with the number of antennas, we consider large antenna deployments also at the BS, with $N = \{1, 4, 8, 64\}$. In this case, the performance of the different precoding strategies is shown in Fig. 4b under the assumption of $M = 40$ blended antennas at the user. Note that in this case, going from 4 to 64 antennas at the BS, we achieve a gain of up to 14× in spectral efficiency, confirming the expected multiplicative antenna gains.

Both Fig. 4a and Fig. 4b clearly show that, in all the simulated regimes, *Optimal S-CSI* and *No CSI* differ in at most 1 bit/s/Hz or 1 dB of SNR. The reason is that, due to the low mutual coupling of the antenna design described earlier, the elements of the channel matrix have very low correlation, i.e., $\mathbb{E}[\mathbf{H}^{\dagger}\mathbf{H}] \approx \mathbf{I}_M$. It is worth emphasizing that the low correlation across the channel matrix elements achieved thanks to the careful antenna design, is a highly desirable feature that allows the MIMOmat-enabled device to fully exploit the multiplicative antenna gains of massive MIMO systems. Note that the performance can be further improved if in the MIMOmat design we allow higher complexity to exploit more accurate CSI. Specifically, assuming perfect instantaneous CSI, either *MF* or *Optimal I-CSI* precoders could be implemented. Observe that the difference in performance between *MF* and *Optimal I-CSI* is small, suggesting that *MF* represents a good compromise between performance and complexity.

Working in Realistic Interference Cellular Scenarios

In order to assess the performance of MIMOmat in a realistic SNR scenario, we now simulate the average uplink achievable throughput in a cellular network with a transmission bandwidth of 5 MHz, in the presence of other interfering users. In the simulations, we assume uniform user distribution and a network consisting of 19 3-sector base stations with N antennas per sector, serving one user per sector at any time-frequency slot. Users are assigned to the BS with least propagation loss [2]. The communication scheme works in TDD mode. In these systems, channel reciprocity can be used to train on reverse link and obtain an estimate of the channel at the transmitter (base station or user, depending on whether downlink or uplink is considered). Specifically, here we assume an uplink consisting of two phases: uplink training and data transmission. The uplink training phase consists of users transmitting training pilots, and base stations obtaining channel estimates. In order to characterize the achievable rates under imperfect channel estimation and pilot contamination, we follow the approach of [14, 15], which assume a MMSE estimator for the channel matrix and provide a lower bound on the achievable rate. According to simulations that are not shown here due to space limitations, the trends in Figs. 4a and 4b still hold in this interference scenario with imperfect CSI, with the MF precoder yielding the best compromise. As such, in this study we consider the MF precoder with the imperfect instantaneous CSI given by the MMSE estimator. Even though the designed antenna array from earlier may use a bandwidth of up to 70–80 MHz, here we choose a bandwidth of 5 MHz to show that even with the limited bandwidth of today's cellular networks, very high data rates can be

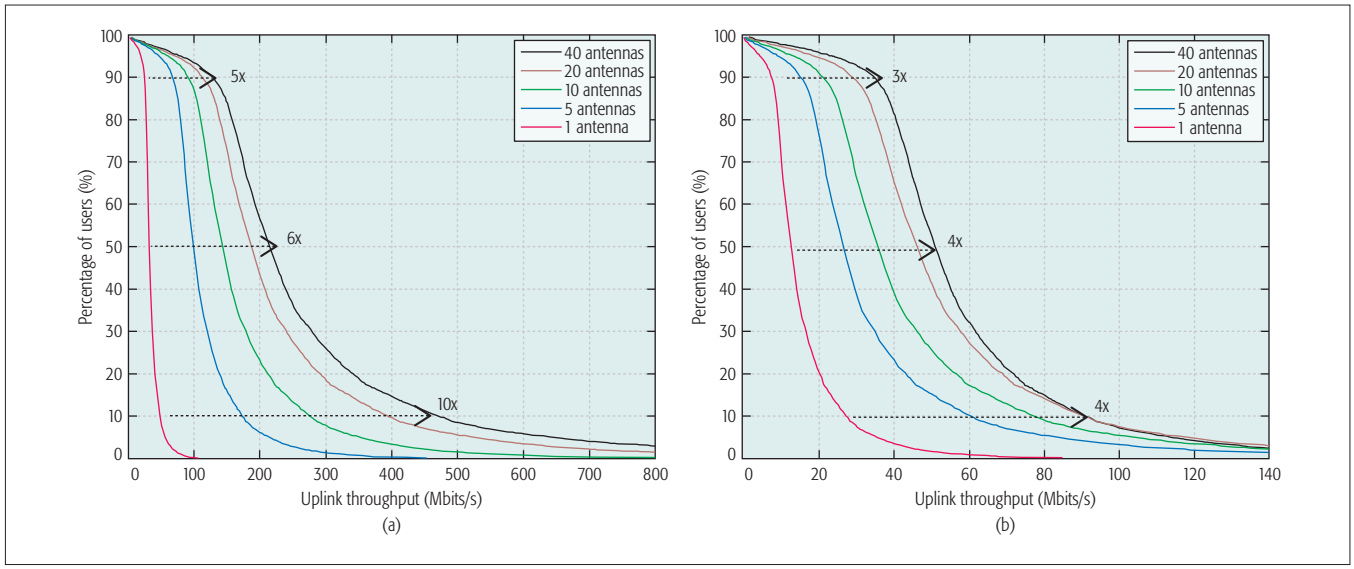


Figure 5. Percentile of users achieving a certain uplink throughput with a signaling bandwidth of 5MHz in a realistic celular scenario [2] when the number of textile antennas range $M = \{1, 5, 10, 20, 40\}$: a) base station with an antenna deployment of $N = 64$; b) base station with an antenna deployment of $N = 4$.

achieved. In this study, we present the cumulative density function (CDF) of the throughput averaged over the fast fading channel realization as a function of the random user locations, for a given transmitted SNR.

In Figs. 5a and 5b, we provide the CDF of the achievable rates for $N = 4$ antennas at the BS (current deployment), and $N = 64$ (future deployment). Each user is assumed to be equipped with $M = \{1, 5, 10, 20, 40\}$ blended antennas. From these two scenarios we can observe dramatic gains as the number of blended antennas goes from 1 to 40 antennas. With an antenna deployment in the BS of $N = 64$ (Fig. 5a), we can get up to 5× in edge rate (covering 90 percent of the users), up to 6× in median rate, and 470 Mb/s in peak rate (10 percent of the users) yielding a gain of 10×. Similarly, at a lower scale, with $N = 4$ antennas at the BS (Fig. 5b), we can get up to 3× in edge rate (covering 90 percent of the users), up to 4× in median rate, and around 50 Mb/s in peak rate (10 percent of the users) yielding a gain of 4×.

Feasibility of a Seamless Lightweight Implementation

Any technological gadget aimed at being seamlessly embedded into the user's personal sphere needs to have very strong requirements in terms of wearability. Ideally, we should move from something portable, where the user would notice that it is carrying a new element, to a solution practically unnoticeable. In this case, the hardware design is key, and aspects such as powering, RF chain design, and the electronics would require further in-depth study. We next comment on each of these issues, detailing the different alternatives that we are currently taking into consideration as ongoing work.

Powering

How the antenna system and its associated signal processing components shall be powered is a strategic decision within the wearable solution design. In the case that the MIMOmat is connected to the user device via a wired interface, a straightforward powering solution is to use the batteries of the device itsetlf. However, while the MIMOmat design is driven by low power requirements, this solution would inevitably lead to the need for additional batteries supplementing the terminal's own, in order to avoid quick battery depletions that would make the whole solution unusable. If the final powering solution is additional batteries, they should be lightweight, high capacity, and obviously rechargeable [11].

RF Chains Design

Conventional massive MIMO systems deploy a single or, at best, very few antennas at the user end [16]. Each antenna typically has one associated RF chain, which is one of the reasons why the number of antennas at the user end is very limited. Figure 6 depicts a traditional multi-antenna design, illustrating the high level of hardware complexity. Clearly, one of the major challenges to make MIMOmat a reality is precisely the RF chains design, which also adds to the powering challenge previously mentioned.

Electronics

As an initial approach, and in order to provide a plug-&-play solution that overrides the device PHY layer, we would need to implement specific signal processing and PHY layer signaling in the MIMOmat. The required electronics are again constrained by space, power consumption, and wearability requirements. One alternative is to implement all the electronics in a unique board, detachable from the textile antenna gadget, that could be charged in a wall socket similarly to a conventional mobile/portable device. Less rigid alternatives are also available, such as the so-called *electronic textiles* that aim at being truly wearable [17], providing useful functionality, while discretely "disappearing" in the fabric.

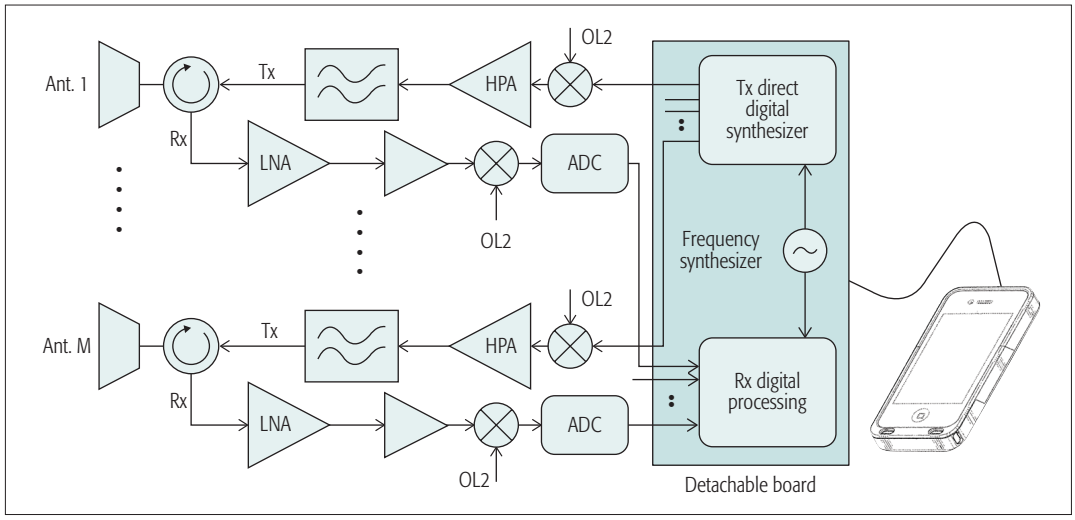


Figure 6. RF chains and processing board sketch.

Conclusions and Future Work

The ultimate success of MIMOmat, a massive MIMO blended antenna wearable extension for the user device, will strongly depend on its seamless usability and affordable cost. While, based on the preliminary design presented in this work, we are confident that MIMOmat will deliver remarkable gains in symmetric data rates, important design aspects need to be addressed, especially those that may jeopardize the wearability, usability, and cost of the solution.

Our efforts are now focused on the hardware implementation, already highlighted earlier, and other issues such as synchronization, duplexing schemes for uplink-downlink communication, channel estimation issues, and sensitivity of channel estimation to signal processing and physical layer signaling. In our view, excluding the hardware implementation, the remaining aspects do not appear critical under the design premises that we manage, and we are therefore confident that they will not significantly compromise the feasibility of our solution, as we expect to show in our future work.

In our view, excluding the hardware implementation, the remaining aspects do not appear critical under the design premises that we manage, and we are therefore confident that they will not significantly compromise the feasibility of our solution, as we expect to show in our future work.

References

[1] F. Boccardi *et al.*, "Five Disruptive Technology Directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 74–80.
[2] H. Huang, C. Papadias, and S. Venkatesan, *MIMO Communication for Cellular Networks*, Springer New York, 2012.
[3] GreenTouch Application Taxonomy Project, "GreenTouch Application Taxonomy," tech. report, GreenTouch, Feb. 2012.
[4] Q. Li *et al.*, "MIMO Techniques in WiMAX and LTE: A Feature Overview," *IEEE Commun. Mag.*, vol. 48, no. 5, May 2010, pp. 86–92.
[5] C. Papadias, "On the Spectral Efficiency of Space-Time Spreading Schemes for Multiple Antenna CDMA Systems," *33rd Asilomar Conf. Signals, Systems, and Computers*, vol. 1, Oct. 1999, pp. 639–43.
[6] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Sig. Proc. Mag.*, vol. 30, no. 1, Jan. 2013, pp. 40–60.
[7] P. Nepa and H. Rogier, "Wearable Antennas for Off-Body Radio Links at VHF and UHF Bands: Challenges, the State of the Art, and Future Trends Below 1 GHz," *IEEE Antennas and Propagation Mag.*, vol. 57, Oct. 2015, pp. 30–52.
[8] P. Van Torre *et al.*, "Indoor Off-Body Wireless MIMO Communication with Dual Polarized Textile Antennas," *IEEE Trans. Antennas and Propagation*, vol. 59, Feb. 2011, pp. 631–42.
[9] Y. Ouyang, D. J. Love, and W. J. Chappell, "Body-Worn Distributed MIMO System," *IEEE Trans. Vehic. Tech.*, vol. 58, May 2009, pp. 1752–65.
[10] Radio Electronics, Resources, and Analysis for Electronic Engineers, "LTE Frequency Bands and Spectrum Allocations," http://www.radio-electronics.com/info/cellulartelecomms/lte-long-term-evolution/lte-frequency-spectrum.php, Oct. 2016.
[11] S. Xu *et al.*, "Stretchable Batteries with Self-Similar Serpentine Interconnects and Integrated Wireless Recharging Systems," *Nature Commun.*, vol. 4, Feb. 2013.
[12] D. Chizhik, "Slowing the Time-Fluctuating MIMO Channel by Beam Forming," *IEEE Trans. Wireless Commun.*, vol. 3, Sept. 2004, pp. 1554–65.
[13] A. Tulino, A. Lozano, and S. Verdú, "Capacity-Achieving Input Covariance for Single-User Multi-Antenna Channels," *IEEE Trans. Wireless Commun.*, vol. 5, Mar. 2006, pp. 662–71.
[14] H. Huh, A. Tulino, and G. Caire, "Network MIMO with Linear Zero-Forcing Beamforming: Large System Analysis, Impact of Channel Estimation, and Reduced-Complexity Scheduling," *IEEE Trans. Info. Theory*, vol. 58, Mar. 2012, pp. 2911–34.
[15] J. Jose *et al.*, "Pilot Contamination and Precoding in Multi-Cell TDD Systems," *IEEE Trans. Wireless Commun.*, vol. 10, Mar. 2011, pp. 2640–51.
[16] S. Wang *et al.*, "Energy-Efficient and Low-Complexity Uplink Transceiver for Massive Spatial Modulation MIMO," *IEEE Trans. Vehic. Tech.*, vol. 64, Oct. 2015, pp. 4617–32.
[17] L. Buechley and M. Eisenberg, "The LilyPad Arduino: Toward Wearable Engineering for Everyone," *IEEE Pervasive Computing*, vol. 7, Apr. 2008, pp. 12–15.

Biographies

Matilde Sánchez-Fernández [SM'14] (mati@tsc.uc3m.es) received the M.Sc. degree in telecommunications engineering and the Ph.D. degree from the Universidad Politecnica de Madrid, Madrid, Spain, in 1996 and 2001, respectively. In 2000, she joined the Universidad Carlos III de Madrid, Madrid, where she has been an associate professor since 2009 teaching several undergraduate and graduate courses (M.Sc. and Ph.D.) related to communication theory and digital communications. Previously, she was a telecommunication engineer with Telefonica. She performed several research stays at the Information and Telecommunication Technology Center, The University of Kansas, Lawrence (1998), Bell Laboratories, Crawford Hill, NJ (2003 – 2006, 2015), Centre Tecnologic de Telecomunicacions de Catalunya, Barcelona, Spain (2007), and Princeton University, Princeton, NJ (2011). Her current research interests are signal processing for wireless communications, multiple-input multiple-output techniques, channel modeling in wireless communications and game theory, and machine learning techniques applied to communications. In these fields, she has (co)authored more than 50 contributions to international journals and conferences.

Antonia M. Tulino [S'00, M'03, SM'05, F'13] (a.tulino@nokia-bell-labs.com) received the Ph.D. degree in electrical engineering from Seconda Universita' degli Studi di Napoli, Italy, in 1999. She held research positions at Princeton University, at the Center for Wireless Communications, Oulu, Finland, and at the Universita' degli Studi del Sannio, Benevento, Italy. In 2002 she joined the Faculty of the Universita' degli Studi di Napoli "Federico II," and in 2009 she joined Bell Labs. Since 2011 Dr. Tulino has been a

member of the editorial board of the *IEEE Transactions on Information Theory*. In 2013 she was elevated to IEEE Fellow. She has received several paper awards, including the 2009 Stephen O. Rice Prize in the Field of Communications Theory for the best paper published in the *IEEE Transactions on Communications* in 2008. She has been the principal investigator on several research projects sponsored by the European Union and the Italian National Council, and was selected by the National Academy of Engineering for the Frontiers of Engineering program in 2013. Her research interests lie in the area of communication systems approached with the complementary tools provided by signal processing, information theory, and random matrix theory.

EVA RAJO-IGLESIAS [SM'08] received the M.Sc. degree in telecommunication engineering from the University of Vigo, Vigo, Spain, in 1996, and the Ph.D. degree in telecommunication engineering from the University Carlos III of Madrid, Madrid, Spain, in 2002. She was a teaching assistant at the University Carlos III of Madrid from 1997 to 2001. She joined the Polytechnic University of Cartagena, Cartagena, Spain, as a teaching assistant in 2001. She joined the University Carlos III of Madrid as a visiting lecturer in 2002, and she has been an associate professor with the Department of Signal Theory and Communications since 2004. She visited the Chalmers University of Technology, Gothenburg, Sweden, as a guest researcher from 2004 to 2008, and she has been an affiliate professor with the Antenna Group, Signals and Systems Department, since 2009. She has co-authored more than 50 papers in JCR international journals and more than 100 papers in international conferences. Her current research interests include microstrip patch antennas and arrays, metamaterials, artificial surfaces and periodic structures, gap waveguide technology, MIMO systems, and optimization methods applied to electromagnetism. She was a recipient of the Loughborough Antennas and Propagation Conference Best Paper Award in 2007, the Best Poster Award in the field of metamaterial applications in antennas at the Metamaterials Conference in 2009, the Excellence Award to Young Research Staff at the University Carlos III of Madrid in 2014, and the Third Place Winner of the Bell Labs Prize in 2014. She is an associate editor of *IEEE Antennas and Propagation Magazine* and *IEEE Antennas and Wireless Propagation Letters*.

JAIME LLORCA (jaime.llorca@nokia-bell-labs.com) received the B.E. degree in electrical engineering from the Universitat Politecnica de Catalunya (UPC), Barcelona, Spain, in 2001, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, MD, USA, in 2003 and 2008, respectively. He held a post-doctoral position at the Center for Networking of Infrastructure Sensors (CNIS), College Park, MD, USA, from 2008 to 2010. He joined Nokia Bell Labs in Holmdel, NJ, USA, in 2010, where he is currently a research scientist in the Network Algorithms Department. His research interests include energy efficient networks, distributed cloud networking, content distribution, resource allocation, network information theory, and network optimization. He is a recipient of the 2007 Best Paper Award at the IEEE International Conference on Sensors, Sensor Networks and Information Processing (ISSNIP), the 2016 Best Paper Award at the IEEE International Conference on Communications (ICC), and the 2015 Jimmy H.C. Lin Award for Innovation.

ANA GARCÍA ARMADA [SM] (agarcia@tsc.uc3m.es) received the Ph.D. degree in electrical engineering from the Polytechnical University of Madrid in February 1998. She is currently full professor at the University Carlos III of Madrid, Spain, where she has occupied a variety of management positions. She is leading the Communications Research Group at that university. She has participated and coordinated several national and international research projects related to wireless communications. She is the co-author of eight book chapters on wireless communications and signal processing. She has published more than 100 papers in international journals and conference proceedings and she holds four patents. She has contributed to international organizations such as ITU and ETSI. She serves on the editorial boards of *Physical Communication*, *IET Communications* and *IEEE Communications Letters*. She has served on the TPC of more than 40 conferences. She has received a Young Researchers Excellence Award and an Award to Best Practices in Teaching, both from the University Carlos III of Madrid. Her main interests are multi-carrier and multi-antenna techniques and signal processing applied to wireless communications.

ACCEPTED FROM OPEN CALL

Impact of IEEE 802.15.7 Standard on Visible Light Communications Usage in Automotive Applications

Alin-Mihai Căilean and Mihai Dimian

ABSTRACT

As the demand for wireless communication technologies is exponentially increasing, using the visible light electromagnetic spectral region is a logical option. This article provides a general analysis of the IEEE 802.15.7 standard for optical communications using visible light, focusing on the PHY I, intended for outdoor applications. Moreover, the article discusses how the standard's specifications apply to visible light communication use in vehicular networking applications. The article concentrates on the standard's applicability in this domain, and points out the fact that it has not been widely accepted, as most of the VLC developers are focused on decreasing system complexity and implementation costs, rather than complying with the standard's requirements. Furthermore, the article points out the necessity of a novel vehicular applications VLC specific standard, which should focus on the requirements imposed in this specific area.

INTRODUCTION

Visible light communications (VLC) represents the use of the visible light (VL) spectrum, between 380 nm and 780 nm, to enable wireless data transfer. The data transfer is achieved as an additional function, besides lighting or signaling. In VLC, the data is modulated onto the instantaneous power of the light and in its simplest form is referred to as on-off-keying (OOK). At the receiver side, the data is extracted using light sensing elements, which can be either photodetector elements or cameras. In the first case, the photodetector, commonly a reversed bias photodiode, is connected in a transimpedance circuit, providing an electrical signal proportional to the power of the incident light. In some cases, the photodetector element is replaced by camera systems [1]. In such instances, the data is extracted using complex image processing techniques, whereas the system performances are strongly related to the ones of the camera.

VLC is an emerging technology that comes with plenty of benefits: financial, technical, medical, and social. Unlike radio frequency (RF) communications, which are classified by the World Health Organization (WHO) as a possible cause of cancer in humans, VLC is safe for humans. Also, VLC is safe for all electronic equipment, enabling it to be used in RF restricted areas (e.g. airplanes, chemical plants or hospitals). Since VLC is based on the existing lighting infrastructure, the technology has a ubiquitous character and a low implementation cost. Furthermore, as the data transfer is achieved as an additional function to lighting or signaling, no extra power is needed for generating the data carrier, and therefore, VLC is an eco-friendly technology. Moreover, as the demand for wireless communications is growing, whereas the RF spectrum is saturating, VLC provides worldwide, unregulated and almost unlimited bandwidth. Therefore, even if it is in its early-stages, VLC can enable short range multi-Gb/s data rates [2].

Inaddition to the high data rate indoor applications, VLC has also been found suitable for transportation safety applications, as in vehicle to vehicle (V2V) [3, 4] and infrastructure to vehicle (I2V) communications [1, 5 6]. In the context of an increasing number of road fatalities (according to the WHO), vehicular communications have the potential to significantly enhance traffic safety. By combining the data collected from neighboring vehicles and from traffic infrastructures, vehicle awareness is considerably increased, as illustrated in Fig. 1. As communication-based vehicle applications are still in their early stages, the requirements for vehicular communication scenarios are rather limited and not broadly accepted by the community. However, several aspects are widely considered as important, such as the high packet delivery ratio (PDR) and the low latencies (below 100 ms). Concerning the data rates and the communication distances, VLC needs to compete with 5.9 GHz dedicated short range communications (DSRC), which aims to achieve distances up to 1000 meters with data rates between 3 Mb/s and 27 Mb/s. However, as summarized in [7], numerous papers have showed that DSRC systems are not able to fully comply with the imposed requirements.

In vehicular applications, VLC is well-suited for heavy traffic situations, such as crowded cities or highways, where due to the numerous neighboring nodes, RF-based communications can be affected by severe packet collisions that increase delays and reduce communication reliability. VLC use does not exclude RF, as the two technologies

The authors provide a general analysis of the IEEE 802.15.7 standard for optical communications using visible light, focusing on the PHY I, intended for outdoor applications. Moreover, they discuss how the standard specifications apply for the visible light communication usage in vehicular networking applications.

Alin-Mihai Căilean is with Stefan cel Mare University of Suceava; Mihai Dimian is with Howard University and Stefan cel Mare University.

Digital Object Identifier:
10.1109/MCOM.2017.1600206

can also be used as complementary solutions. Actually, VLC is appropriate for high traffic densities, and therefore short distances, whereas 5.9 GHz DSRC is suitable for long distances [7]. Furthermore, VLC could be used to take some of the load off the RF network, in order to improve its performance.

Since communication-based vehicle applications are considered to be the next generation of vehicle safety systems, the use of VLC in this domain is quite logical. As the performance of VLC technology has been confirmed with its standardization by IEEE, this article provides an overview of the IEEE 802.15.7 standard for optical communications using VL. However, unlike other papers that discussed this issue (e.g. [8]), *this work is focused on the PHY I, dedicated to outdoor low data rate applications, and aims at determining how the standard requirements and specifications comply with VLC use in automotive applications*. Furthermore, this article proposes several amendments that could further enhance the compliance to vehicular applications.

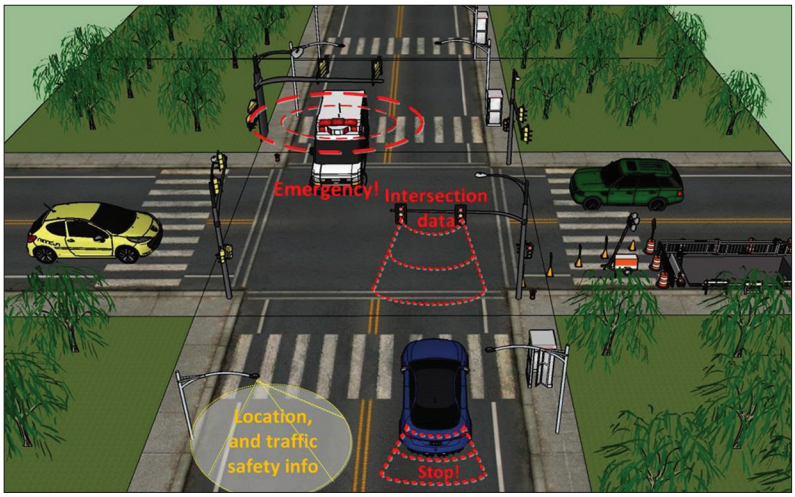


Figure 1. VLC usage scenario: road safety data is transmitted using the vehicle lighting systems, the street lighting system and the traffic lights.

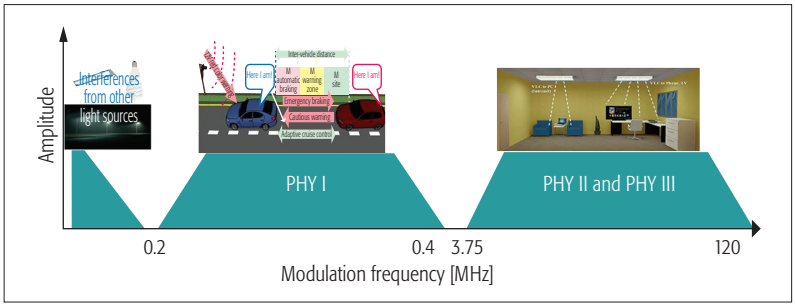


Figure 2. Frequency division multiplexing for the three PHY types.

	Infrastructure	Mobile	Vehicle
Fixed coordinator	Yes	No	No
Power supply	Ample	Limited	Moderate
Form factor	Unconstrained	Constrained	Unconstrained
Light source	Intense	Weak	Intense
Physical mobility	No	Yes	Yes
Range	Short/long	Short	Long
Data rates	High/Low	High	Low

Table 1. Device classification according to IEEE 802.15.7 [9].

Considerations on the IEEE 802.15.7 Standard and Its Effect on Vehicular Applications

The IEEE 802.15.7 standard [9] for short-range wireless optical communication using VL was released in September 2011. The current version of the standard covers the physical layer (PHY) and the medium-access control (MAC). According to the standard, the wireless data transfer is accomplished by modulating the intensity of optical devices, such as LEDs, at frequencies imperceptible to the human eye, and without affecting in any way the primary role of the device, which is lighting or signaling. Therefore, the aspects associated with flickering and high resolution dimming are seriously taken into consideration. The standard also considers the issues regarding link mobility, the impairments caused by noise, and the interference from other light sources.

Depending on the applications and the required data rates, the IEEE 802.15.7 standard comes with three PHY types. PHY I is envisioned for outdoor low data rate applications, and uses OOK and variable pulse position modulation (VPPM), with data rates between 11.67 kb/s and 267 kb/s. PHY II and PHY III are proposed for indoor moderate data rate applications, with data rates between 1.25 Mb/s and 96 Mb/s. PHY II also uses OOK and VPPM, whereas PHY III is intended for color-shift-keying (CSK) applications. CSK is achieved using multicolor LEDs and color selective photodetectors. In this case, the VL spectrum is divided in seven bands, and the data is encoded by using variable combinations of three colors, according to a mapping rule. The three physical layers can coexist but cannot interoperate. As illustrated in Fig. 2, PHY I is situated on a spectral region different from the region for PHY II or PHY III, enabling frequency division multiplexing (FDM) as a coexistence technique.

The following subsections aim at highlighting how VLC automotive applications cope with standard specifications. They address the classes of VLC devices, the medium access control (MAC) topologies, aspects related to dimming and flickering, to modulation and data rate, as well as to the structure of the data frame.

Classes of VLC Devices

Referring to the VLC devices, the standard specifies three different classes. Table 1 summarizes the particular features for each class. A gratifying aspect is that the standard mentions here the automotive domain, as a possible VLC application area. Although briefly, it defines several aspects associated with networking and the imposed requirements.

MAC Topologies

As shown in Fig. 3, the IEEE 802.15.7 MAC supports three access topologies: star, peer-to-peer, and broadcast. The identification of the VLC

devices is done by using a 64 bit address, whereas when a device becomes a coordinator, it will be identified with a 16 bit address. The star topology involves the presence of a coordinator, which is responsible for the control of the communication. In this case, the devices will form an independent network, with a unique VLC personal area network (VPAN) identifier.

The communication is possible only between devices that joined the network, whereas other devices can join after the coordinator allows them. On the other hand, the peer-to-peer topology is significantly more permissive, with every device being able to communicate directly to all the devices within its vicinity. Therefore, one of the two nodes involved, usually the node that initiates the communication, acts as coordinator. The third topology is the broadcast topology, which implies data transmissions from one node to another, or to more nodes, without forming a network. This type of communication is unidirectional and no destination address is required.

In vehicle safety applications, the central objective is to transmit information concerning the vehicle position (e.g. GPS coordinates) and its state (e.g. velocity, acceleration, direction) toward the neighboring vehicles. In addition to these routine messages, another foremost message category contains the event driven messages. These messages are generated by an unexpected change in behavior or by a potentially dangerous situation, and are granted with the highest priority.

A crucial aspect in vehicular communications is the fact that in this environment, the data is location distributed and not individually addressed. A moving vehicle is continuously transmitting geographically distributed data within its surroundings (in front, behind, and lateral). In such a scenario, *the channel access should be without a coordinator*. Furthermore, in most instances, the applications involve single-hop communications, and therefore, there is no networking required. Another particularity of vehicle ad-hoc networks (VANETs) comes from their highly dynamic character. As the vehicles are continuously moving, the data is continuously updated (new location, new state), resulting in an increased message generation rate (at least 10 messages/second). In this case, the vehicle transmitting a message is expecting that the receiving vehicles will take adequate responses that reduce the danger, rather than to respond with an acknowledgment.

In light of the above-mentioned, one can see that in a first step, vehicular communication applications will be mainly based on the broadcast topology, as most of *the data transmissions are geographically distributed and have no destination address*. However, in future applications, as in platooning, where all the participants respond to the received messages by transmitting their own location and status, the communication will be mainly peer-to-peer. However, non-line-of-sight (LoS) conditions (intersection crossing assistance applications or multi-hop scenarios) are examples of situations where a coordinator-based star topology is better suited. In such cases, a central node (e.g. the traffic light) will facilitate the communications between nodes. Therefore, *due to the high dynamicity of the network and depending on the occurring events, vehicular networks are developing toward mesh topology networks.* This will involve the necessity of complex routing protocols, which must provide robust connectivity and rapid access times. *Thus, a future version of a vehicular VLC standard should be developed considering different types of (safety) applications. This version should cover in detail the routing protocols that should be selected depending on the type and on the priority of the messages.*

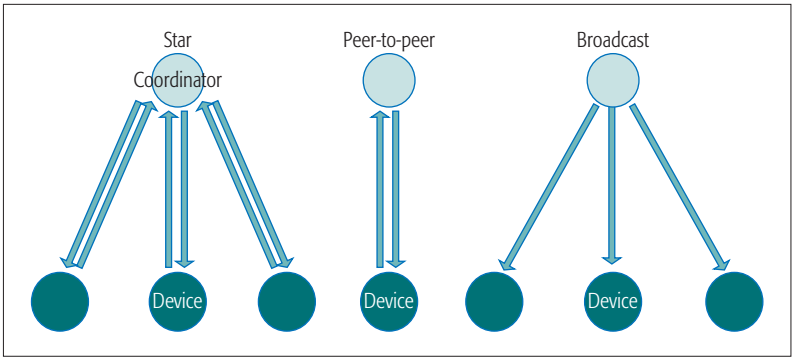


Figure 3. IEEE 802.15.7 standard visible light communication medium access topologies.

Dimming

A specific problem in VLC compared to other wireless communication technologies is the necessity to maintain data transfer while radiation is dimming. The standard proposes two different dimming solutions. In OOK, the light intensity levels of the *on* and *off* states remain unchanged, whereas the dimming is achieved through the insertion of compensation times. The compensation times represent additional time slots in which the light is completely turned on or off, depending on the required dimming level. Although the maximum communication distance remains constant in this case, the insertion of compensation times affects the transfer data rate. Furthermore, it also leads to a loss of synchronization, which imposes the insertion of additional short resynchronization patterns.

In VPPM, dimming is achieved by controlling the pulse width, and thus by modifying the bits, controlling light intensity levels. Let us remember that VPPM combines the features of pulse position modulation (2-PPM) to prevent flickering and pulse width modulation (PWM) to enable dimming and brightness control. As opposed to OOK, the maximum communication distance is reduced in this case while the data rate remains constant. VPPM is mainly envisioned for applications that involve high resolution dimming and where the data rate is prior to the communication distance. In transportation, VPPM seems to be suitable for energy saving applications based on light intensity control, as in intelligent street lighting systems, which can be further enhanced with data broadcasting capabilities (I2V). Here, the distances between the street light and the vehicles on the street are constant, predictable, and relatively short. Therefore, even with the lights dimmed, high data rates are enabled, whereas the short distances can still be safely covered. On the other hand, vehicular VLC applications also imply scenarios in which the communication distance and the connectivity are more important than the

Modulation	RLL coding	Optical clock	FEC		
			Outer code (RS)	Inner code (CC)	Data rate ([kb/s)
OOK	Manchester	200 kHz	(15,7)	1/4	11.67
			(15,11)	1/3	24.44
			(15,11)	2/3	48.89
			(15,11)	None	73.3
			None	None	100
VPPM	4B6B	400 kHz	(15,2)	None	35.56
			(15,4)	None	71.11
			(15,7)	None	124.4
			None	None	266.6

Table 2. PHY I operating modes [9].

data rate. The communication between a traffic light and the approaching vehicles, or the communication between two vehicles, are relevant examples for such applications. In such cases, OOK seems to be more appropriate.

The fact that the standard has considered these two different situations and is offering adequate dimming options for both circumstances can only be appreciated, as it enables VLC use in multiple scenarios.

Flickering

Another VLC-specific problem is related to the light flickering that might be induced by the modulation technique. Different from other wireless communication technologies, in VLC, the light wave carrier is perceivable by the human eye. Consequently, the standard strictly imposes that the modulation method must not induce any noticeable flickering that might affect human health. The standard defines flicker as a brightness fluctuation that can produce noticeable physiological changes in humans. In consequence, it delimits the maximum flickering time period (5 ms), within which the light intensity can be changed without being perceived by the human eye [9].

In VLC, flickering is categorized as intra-frame flickering (within the data frame) and as inter-frame flickering (between adjacent frames). To prevent inter-frame flickering, the standard stipulates the use of idle patterns that have brightness equal to one of the data frames. Intra-frame flickering is prevented by using run length limited (RLL) coding. In RLL coding, an equal number of ones and zeros are generated and consequently, the flickering is diminished by avoiding long series of ones or zeros. For example, in outdoor applications, the standard specifies the use of Manchester coding for OOK, and 4B6B coding for VPPM.

Although the standard seems to be very strict concerning flickering, the effect of the VLC induced flickering is rather insufficiently studied. The concerns related to possible health issues generated by VLC flickering might justify such strict constraints, but they could be different from one class of applications to another. On the one hand, indoor applications are associated with a long exposure time and require more care in addressing the flickering issue. On the other hand, in vehicular applications, the exposure to the communication light is less direct, less intense, and the potential flickering source is less often the main lighting source. Furthermore, as the vehicle is moving, the exposure to communication light is limited in time.

Based on these considerations, unless clear evidence of health problems caused by modulation induced flickering are found, *the standard could be less strict to flickering for vehicular applications. In this direction, the standard could allow the use of other codes and/or modulations under some specific constraints.*

Modulation Frequencies, Forward Error Correction, and Data Rates

Unlike the indoor applications, for which the standard stipulates the use of 3.75 MHz to 120 MHz optical clock rates, for outdoor applications, optical clock rates below 400 kHz were chosen. Lower frequencies were considered because LEDs used in outdoor applications (e.g. street lighting, traffic lights) generally require high currents and therefore they switch slower. For OOK applications, the standard specifies the use of a 200 kHz optical clock, whereas for VPPM, an optical clock rate of 400 kHz is specified. These clock rates were considered in order to prevent any interference with other sources of light, which may generate harmonics with frequencies of up to a few tens of kilohertz, but also to prevent flickering.

As the outdoor VLC applications involve longer distances, these VLC links are strongly affected by path loss. Furthermore, outdoor applications are disrupted by the strong daylight interference and also by other artificial light sources. To mitigate the effect of the unfriendly conditions, the standard specifies the use of convolutional codes (CC) in addition to the Reed Solomon (RS) codes, specified for the indoor applications. The RS and the CC blocks are separated by an interleaver, providing a 1 dB performance improvement. Furthermore, the two FEC codes are highly compatible with the RLL codes, which also have error detection capabilities, insuring another 1 dB improvement.

As summarized in Table 2, although significantly improving the robustness to noise, the FEC codes significantly affect communication data rates and throughput. Furthermore, their use generates an increased number of computations, and therefore they entail the use of a powerful and more expensive data processing unit. *As the automotive industry is rather budget-restrictive, a fair cost-performance tradeoff should be sought.*

Frame Structure

Concerning the data frame structure, as illustrated in Fig. 4, the standard proposes a rather minimalist frame, comprised of three main fields: synchronization header, physical header, and data field.

The frame begins with a synchronization locking pattern, which enables the receiver to achieve optical clock synchronization. Next, the preamble includes a sequence of four topology-dependent

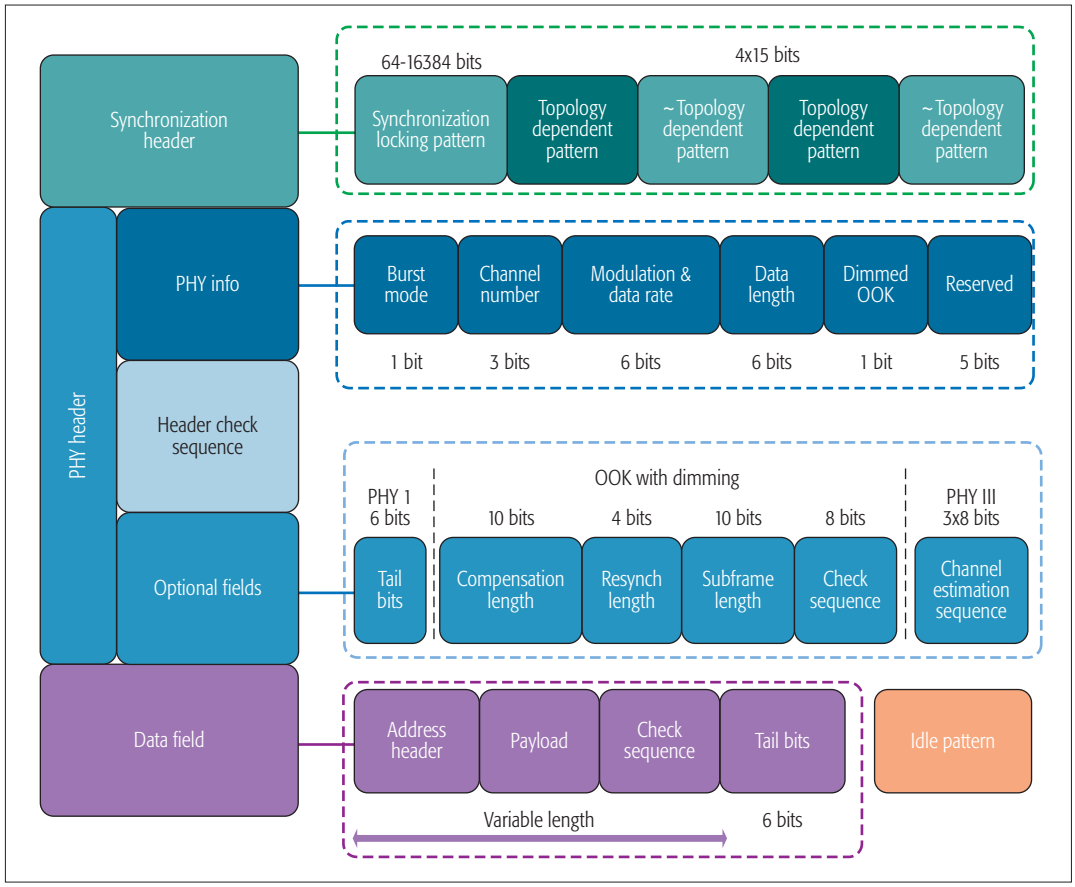


Figure 4. Structure of the VLC data frame.

patterns which provide information concerning the selected topology (i.e. broadcast, star, or peer-to-peer). The preamble field is transmitted using OOK modulation and no channel coding.

The second main field is the PHY header, providing information regarding the transmission mode, the number of the communication channel, the selected PHY layer, the data rate, and the message length. As this information is vital for the proper decoding of the data, the header is protected by a 16 bit cyclic redundancy check sequence. The next field of the PHY header encloses some optional fields, specific for particular situations. The transmission of the PHY header is performed at the lowest data rate by using OOK and Manchester encoding.

The third main field is the data field, which has a length of up to 1023 bytes and is protected by a frame check sequence. Idle patterns are also introduced to separate the data frames in order to prevent inter-frame flickering.

By taking into account the specific characteristics of vehicular communications, such as short data messages (200 bits to 1000 bits), the need for very low delays and high PDR, and the lower requirements for dimming, *the standard could be adapted by reducing the overhead*. For example, the synchronization header and optional fields could be reduced. Special attention should be paid to the check sequences, with the length resulting from a compromise between BER and throughput, and therefore, it can be determined by the requirements of the vehicular applications, and not by overall VLC requirements.

Discussions on Vehicular VLC Standardization

Benefits and Challenges of a Vehicular VLC Standard

In the last decade, the performance of LED devices has significantly improved, while the cost been decreasing. Within this context, LED light sources are meant to totally replace classical lighting. This trend is also observable in transportation, where LEDs are used in traffic lights, street lighting systems, traffic displays, or in vehicle lighting systems. Referring to transportation, this great dispersal offers the premises for an unprecedented distribution of the VLC technology, which in turn can ensure fast technology deployment and a rapid market penetration. However, an important step toward technology deployment is standardization. Standardization of the work has the potential to diminish the gap between academia and industry and also to stimulate the technology. Therefore, a competitive and widely accepted standard offers great benefits for both industry and consumers.

An important step toward technology deployment is standardization. Standardization of the work has the potential to diminish the gap between academia and industry and also to stimulate the technology. Therefore, a competitive and widely accepted standard offers great benefits for both industry and consumers.

Limited Acceptability of the IEEE 802.15.7 Standard

Despite the fact that the IEEE 802.15.7 standard was published several years ago, one can observe that the standard has not been widely accepted. Currently, there are only a few works that report hardware implementations of VLC systems that comply with the standard's requirements (e.g. [6, 10–12]). The divergence between the state of the art and the standard is most obvious in the

A gratifying aspect concerning the future development of the VLC technology is the fact that a revision of the standard, known as IEEE 802.15.7r1, is already under development. This revised version includes vehicular communications as a fundamental VLC use case, mentioning here V2I and V2V applications.

case of indoor VLC systems. Here, the standard stipulates data rates up to 96 Mb/s, in the context of numerous experiments that proved data rates above 1 Gb/s [2]. Although multi-Gb/s data rates are often reported for centimeter distances, there are several prototypes that achieved Gb/s data rates up to a few meters [2]. Furthermore, a significant number of the VLC developers are considering using other modulations and coding techniques rather than those of the standard.

Referring to its application in V2V or in I2V communication, again, the standard's use is limited. Moreover, *most of the existing VLC systems targeted for inter-vehicular communications have been developed without even attenpting to comply with the standard's requirements [1].* In other cases [4, 5], the developers used the specified modulation and/or coding, but without complying with the other architectural or functional requirements (e.g. frame structure, modulation frequency, data rates).

Generally, prototype developers have chosen to use solutions different from those of the standard with the purpose of maintaining low system complexity, and thus a low price, factors that are crucial for the automotive industry. In [6], the authors present the experimental results of one of the very few standard-complying VLC receivers intended for automotive applications. Although a low cost was the goal, the prototype had a final price of more than 1500€. One can expect that once widely accepted, mass production will reduce the cost, but still, such a price level is too high. Furthermore, due to the complexity of FEC protocols (complexity in terms of high computing times), the system was not able to process the data in real time, imposing the use of more complex data processing units. Concerning the throughput efficiency, it is below 50 percent, whereas the BER results were between 10^{-5} and 10^{-3}, for distances up to 10 meters. In these circumstances, the results appear to be rather unsatisfactory. The low throughput efficiency of standard-complying VLC systems is also confirmed in [10], whereas in [11] and [12], the proposed systems achieve communication distances of only a few meters, too short for practical use in vehicular applications.

Automotive-Specific Concerns and Issues That Should be Covered by a Future Vehicular VLC Oriented Standard

Although lapidary, the inclusion of intelligent transportation system (ITS) applications in the IEEE 802.15.7 standard is encouraging the continued research in this domain. However, even though the standard refers to the outdoor low data rate applications and mentions here vehicle communications, it was not specially developed for such a purpose. Furthermore, the standard does not provide the required specifications that should be applied in the field, offering only a general framework. Therefore, even though the standard identifies transportation domain as a potential application area, it does not provide any specific information regarding I2V or V2V implementing rules. Thus, developers in the field do not refer to the standard when developing their prototypes. However, starting from the current specifications of the standard, after identifying the vehicular communication-specific requirements, a new standard release could be expected. A similar approach was considered for 5.9 GHz RF communications, where the IEEE 802.11p standard for wireless access in vehicular environments was developed as an extension of the IEEE 802.11a standard for wireless local area networks (WLAN).

Vehicular safety applications must provide highly reliable links and are very stringent in terms of latencies and PDRs. Within this context, the VLC systems designed for automotive applications must have the ability to cope with the vehicular environment (i.e. highly dynamic and very unpredictable) and with the specific atmospheric conditions (i.e. strong sun light, fog, rain or snow). Supporting reliable and robust communications in such conditions represents a major challenge for VLC, whereas having doubts on these issues further postpones technology deployment. Part of the problem could be addressed by introducing environment-adaptive communications [13]. Thus, the standard should provide a channel estimation algorithm, enabling the vehicles to adjust the communication parameters (e.g. modulation, data rate) depending on the external conditions and on the SNR level. Moreover, the standard should allow data ranking according to the priority of the information, along with different quality of service (QoS) requirements for each case (i.e. priority class). Accordingly, the FEC usage should also be priority oriented, and thus the messages should also be classified according to their relevancy (i.e. messages relevant only for the neighboring vehicles, for a wider area, or messages that require retransmission and a specific time to live period).

Therefore, the VLC standard for vehicular applications should be very strict concerning the robustness to perturbations and to latencies, whereas it could be less stringent concerning flickering mitigation and high resolution light dimming, which are less important in this area. Furthermore, the future standard should further simplify the frame structure, in order to reduce the overhead and enhance the throughput. Other aspects that should be approached are related to the message generation rate and to vehicular-specific networking. Here, the standard should provide information referring to dynamic mesh topologies, able to provide rapid and efficient channel access.

In addition to lighting and wireless communications, VLC could also be used in distance measurement and positioning applications, as in [3, 14]. Although in an early stage, the potential of this type of application is high, whereas the benefits in vehicular safety applications are even higher. Thus, as the technical aspects are clarified and the viability experimentally demonstrated, a future revision of the standard should cover the aspects regarding the VLC usage for inter-vehicle distance measurements and positioning.

A gratifying aspect concerning the future development of the VLC technology is the fact that a revision of the standard, known as IEEE 802.15.7r1, is already under development. This revised version includes vehicular communications as a fundamental VLC use case, mentioning here V2I and V2V applications. In this case, the standard considers the requirements of vehicular communications and aims to enhance mobility, data rates, robustness, and to enhance the networking protocols [15].

Conclusions

As the interest in communication-based vehicle safety applications is increasing, VLC provides an efficient solution for dense traffic situations. This article has reviewed the IEEE 802.15.7 standard for optical communications and debated the aspects related to its application in automotive communications.

The article has pointed out that although it has been several years since it was released, the standard acceptance is quite limited, as the developers in the field prefer to develop their prototypes independent from the standard rather than complying with its requirements. Currently, it can be considered that complying with the standard's specifications increases the system's complexity and the implementation cost. Furthermore, as automotive applications generally use short messages, the throughput performances are significantly affected by the large overhead. Within this context, the article has pointed out the necessity of a standard focusing on the usage of VLC in vehicular applications. An intermediate step toward this objective could be represented by the inclusion of VLC vehicular communication as a use case of the IEEE 802.15.7r1 standard.

Acknowledgment

The infrastructure used in this work was partially supported by the project "Integrated Center for Research, Development and Innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for Fabrication and Control," contract No. 671/09.04.2015, Sectoral Operational Program for Increase of the Economic Competitiveness co-funded by the European Regional Development Fund. This work was partially supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI - UEFISCDI, project number PN-III-P2-2.1-PED-2016-2011, within PNCDI III.

References

[1] T. Yamazato, *et al.*, "Image-Sensor-Based Visible Light Communication for Automotive Applications," *IEEE Commun. Mag.*, vol. 52, no. 7, July 2014, pp. 88–97.

[2] Y. Wang, *et al.*, "8-Gb/s RGBY LED-Based WDM VLC System Employing High-Order CAP Modulation and Hybrid Post Equalizer," *IEEE Photonics J.*, vol. 7, no. 6, Dec. 2015, pp. 1–7.

[3] S.-H. Yu, *et al.*, "Smart Automotive Lighting for Vehicle Safety," *IEEE Commun. Mag.*, vol. 51, no. 12, Dec. 2013, pp. 50–59.

[4] A.-M. Cailean, *et al.*, "Visible Light Communications: Application to Cooperation Between Vehicles and Road Infrastructures," *Proc. IEEE Intell. Vehicles Symp. (IV)*, June 2012, pp. 1055–59.

[5] A.-M. Cailean, *et al.*, "Novel Receiver Sensor for Visible Light Communications in Automotive Applications," *IEEE Sensors J.*, vol. 15, no. 8, Aug. 2015, pp. 4632–39.

[6] A. Belle, *et al.*, "Development of IEEE802.15.7 based ITS Services Using Low Cost Embedded Systems," *Proc. 13th Int. Conf. ITS Telecommunications*, Nov. 2013, pp. 419–25.

[7] A.-M. Cailean, *et al.*, "A Survey on the Usage of DSRC and VLC in Communication-Based Vehicle Safety Applications," *Proc. IEEE 21st Symp. Commun. Vehic. Tech. Benelux (SCVT)*, Nov. 2014, pp. 69–74.

[8] S. Rajagopal, R. D. Roberts, and S.-K. Lim, "IEEE 802.15.7 Visible Light Communication: Modulation Schemes and Dimming Support," *IEEE Commun. Mag.*, vol. 50, no. 3, Mar. 2012, pp. 72–82.

[9] IEEE Standard for Local and Metropolitan Area Networks–Part 15.7: Short-Range Wireless Optical Communication Using Visible Light, IEEE Standard, Sept. 2011, 1-309.

[10] J. Baranda, P. Henarejos, and C. G. Gavrincea, "An SDR Implementation of a Visible Light Communication System based on the IEEE 802.15.7 Standard," *Proc. 20th Int'l. Conf. Telecomm. (ICT)*, May 2013, pp. 1–5.

[11] C. Gavrincea, J. Baranda, and P. Henarejos, "Rapid Prototyping of Standard-Compliant Visible Light Communications System," *IEEE Commun. Mag.*, vol. 52, no. 7, July 2014, pp. 80–87.

[12] J. Higuera, *et al.*, "Tuneable and Portable Lighting System for Visible Light Communications (TP-VLC)," *Proc. IEEE Int'l. Instrumentation and Measurement Techn. Conf. (I2MTC)*, May 2015, pp. 91–96.

[13] A.-M. Cailean and M. Dimian, "Toward Environmental-Adaptive Visible Light Communications Receivers for Automotive Applications: A Review," *IEEE Sensors J.*, vol. 16, no. 9, May 1, 2016, pp. 2803–11.

[14] T. Yamazato, *et al.*, "Image Sensor Based Visible Light Communication and Its Application to Pose, Position, and Range Estimations," *IEICE Trans. Commun.*, vol. E97-B, no. 9, 2014, pp. 1759–65.

[15] V. Jungnickel *et al.*, "A European View on the Next Generation Optical Wireless Communication Standard," *Proc. IEEE Conf. Standards for Commun. Net. (CSCN)*, Tokyo, 2015, pp. 106–11.

Biographies

Alin-Mihai Căilean (alinc@eed.usv.ro) received a B.S. degree in electrical engineering (2009) and a M.S. in computer and communication networks (2011) from the University of Suceava, Romania. He received his Ph.D. (2014) after a joint program between the University of Versailles St. Quentin en Yvelines, France and the University of Suceava. Currently, he is a researcher at the University of Suceava. His research is focused on visible light communications, wireless sensors, and vehicle safety applications.

Mihai Dimian (dimian@usm.ro) received his B.S. in mathematics (1997) and in physics (2001), as well as a M.S. in dynamical systems from the University of Iassy, Romania. He graduated with a Ph.D. in electrical engineering (2005) from the University of Maryland, College Park, MD, USA, and performed post-doctoral research at the Max Planck Institute, Leipzig, Germany. He is an associate professor at Howard University, Washington D.C., USA, and a professor at the University of Suceava, Romania.

Although it has been several years since it was released, the standard's acceptance is quite limited, as the developers in the field prefer to develop their prototypes independent from the standard rather than complying with its requirements.

ACCEPTED FROM OPEN CALL

Graph-based Cyber Security Analysis of State Estimation in Smart Power Grid

Suzhi Bi and Ying Jun (Angela) Zhang

The authors introduce a graph-based framework for performing cyber-security analysis in power system state estimation. Compared to conventional arithmetic-based security analysis, the graphical characterization of state estimation security provides intuitive visualization of some complex problem structures and enables efficient graphical solution algorithms, which are useful for both defending and attacking the ICT system of smart grid.

ABSTRACT

The smart power grid enables intelligent automation at all levels of power system operation, from electricity generation at power plants to power usage in the home. The key enabling factor of an efficient smart grid is its built-in ICT, which monitors the real-time system operating state and makes control decisions accordingly. As an important building block of the ICT system, power system state estimation is of critical importance to maintain normal operation of the smart grid, which, however, is under mounting threat from potential cyber attacks. In this article, we introduce a graph-based framework for performing cyber-security analysis in power system state estimation. Compared to conventional arithmetic-based security analysis, the graphical characterization of state estimation security provides intuitive visualization of some complex problem structures and enables efficient graphical solution algorithms, which are useful for both defending and attacking the ICT system of the smart grid. We also highlight several promising future research directions on graph-based security analysis and its applications in smart power grid.

INTRODUCTION

The smart power grid is committed to providing stable, high-quality and inexpensive electricity supply to meet the surging power demand of modern society through its intelligent energy management in power generation, transportation and distribution, and its introduced competitive market mechanisms. Essentially, the intelligence of smart grid is driven by its embedded ICT infrastructure, especially the EMS/SCADA (energy management system and supervisory control and data acquisition) system [1]. As shown in Fig. 1, the SCADA system is responsible for collecting the measurement data reported by distributed meters/sensors, which is then fed to the state estimator located at the control center to derive the estimation of system state variables, for example, bus voltage amplitudes and phases. Based on the estimation, the EMS, as well as other power system applications, then makes control decisions, for example, optimal power flow, load curtailment, and electricity pricing, to adjust the physical aspects of the power grid. Evidently, a secure and efficient power system requires accurate state estimation that truthfully reflects the system operating state.

The dependence of smart grid on its ICT infrastructure makes cyber-attacks on state estimation a viable approach to impact normal system operation. In the conventional power network, power devices are isolated from the public network and under close control by the industrial system operator. In the smart grid, however, many distributed smart meters are installed in households, which often connect to the public Internet and run IP-based communication protocols to facilitate two-way information exchange between the users and system operator. This computer-network-like ICT structure achieves low management cost, but also exposes the smart grid to potential cyber attacks through the public information access points. One common cyber attack in smart grids is false-data injection, which distorts the measurements collected by the system operator through either physical device compromise or remote cyber-data injection [2]. Being able to compromise the state estimation, an adversary capable of false-data injection can have a large impact on the power system and beyond, such as earning lucrative profits from electricity price manipulation in the power market [3, 4], or causing a regional blackout to induce chaos and financial loss [5].

The state estimator commonly uses a bad data detection (BDD) mechanism to filter faulty data, either caused by random network error or malicious injection [1]. However, BDD is unable to detect some structured collaborating injection attacks that are disguised as normal measurements [2]. One countermeasure is data-driven detection, which uses the statistical features of the previously collected measurement data to identify anomalous measurements [4]. Nonetheless, it cannot fully eliminate the threat of injection attacks, and its performance highly depends on the accuracy of the extracted statistical features. To fundamentally mitigate false-data injection attacks, it is necessary to secure meter measurements themselves to evade malicious injections by, for example, guards, video monitoring, or tamper-proof communication systems [6]. In a large power network with hundreds of meter measurements, it is tempting to devise a *strategic* protection that achieves system security requirements with low cost, for example, a small number of secured devices.

Arithmetic and *graphical* methods are two popular approaches for security analysis in power system state estimation. Specifically, the arithmetic approach applies algebra and matrix theory to analyze the solution space of state estimation,

Digital Object Identifier:
10.1109/MCOM.2017.1600210C

Suzhi Bi is with Shenzhen University; Ying Jun Zhang is with The Chinese University of Hong Kong.

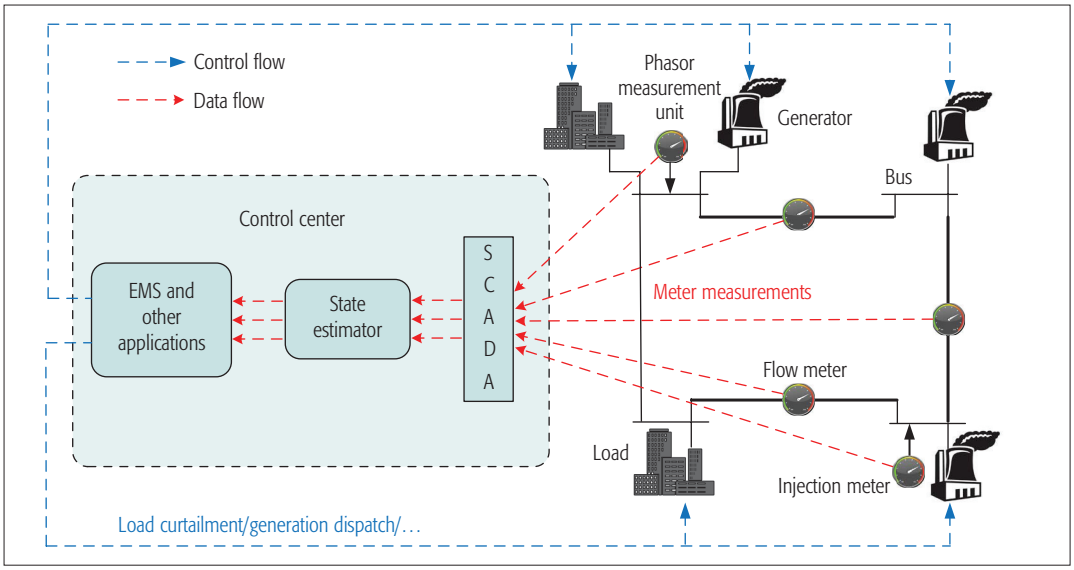


Figure 1. An illustration of the operation of the SCADA/EMS system for a four-bus network.

In recent years, phasor measurement units have emerged as an advanced metering technology that can provide direct real-time voltage phasor measurement with high accuracy and reliability in addition to the conventional meters. In practice, due to high PMU installation costs and the legacy power system in operation, state estimation is often obtained from a mixture of PMUs and power flow measurements.

and thus the potential threats and countermeasures of injection attacks (e.g., [7, 8]). Despite its effectiveness in extensive applications, the arithmetic approach is found to be inefficient in handling some complex problems, especially for those with combinatorial features, for example, involving selecting k out of K buses. Alternatively, the graph-based approach, which uses graph models to characterize security problems, can provide intuitive visualization of complex problem structures (e.g., [9–12]). Its useful insight can lead to efficient optimal or sub-optimal graphical solution algorithms that are otherwise not achievable by arithmetic approaches. However, classic graph algorithms often need significant modifications to solve power system security problems of unique graphical structures.

In this article, we provide an overview of graphical methods for performing cyber-security analysis in power system state estimation. Specifically, we first describe the method to model power network in a graph. Then, we establish a graph-based characterization of state estimation security, and introduce representative graphical algorithms to solve security problems in state estimation. We also suggest several future research directions on graph-based security analysis and its applications in smart power grid. Finally, we conclude this article.

Graph Modeling of Power Networks and Measurements

As shown in Fig. 2a, a power network consists of a number of buses, loads, power generators, and power transmission lines that interconnect them.[1] One important parameter representing the operating state of the power system is the phasor of each bus, including its voltage phase angle and voltage magnitude. In practice, the voltage magnitudes can often be directly measured, while the values of phase angles need to be obtained from state estimation [1]. Conventionally, in the linearized DC measurement model, the estimate of the phase angles is obtained from the active power measurements, that is, the active power flows along the power lines (e.g., meter 1) and the active power injections at the buses (e.g., meter 2). In recent years, phasor measurement units (PMUs) have emerged as an advanced metering technology that can provide direct real-time voltage phasor measurements with high accuracy and reliability in addition to the conventional meters. In practice, due to high PMU installation costs and the legacy power system in operation, state estimation is often obtained from a mixture of PMUs and power flow measurements.

For a power network with $n + 1$ buses, we regard one of them as the reference bus, denoted by R, and estimate the phase angles of the other n buses (state variables) from m meter measurements, denoted by $\theta = (\theta_1, \theta_2, .., \theta_n)'$ and $\mathbf{z} = (z_1, z_2, ..., z_m)'$, respectively. Also, we denote the set of n unknown buses as $\mathcal{S}$, the set of all the buses $\mathcal{V} \triangleq R \cup \mathcal{S}$, the set of transmission lines as $\mathcal{E}$, and the set of m measurements as $\mathcal{M}$.

As shown in Fig. 2b, a power network can also be described in an undirected graph, where vertices and edges represent buses and transmission lines, respectively. Without loss of generality, we regard bus 1 as the reference throughout this article. Loosely speaking, a flow meter reflects the difference between two state variables; an injection meter reflects the sum of differences of a state variable with respect to the subset of state variables in one-hop distance; and a PMU meter reflects the difference of a state variable with respect to the reference bus. For the convenience of exposition, we consider in this article only conventional power flow measurements. In fact, a PMU measurement can be equivalently converted to a flow measurement in security analysis, which is discussed in [9].

Given a subset of meter measurements $\overline{\mathcal{M}} \subseteq \mathcal{M}$, we can find correspondingly a subnetwork (and thus a subgraph) measured by $\overline{\mathcal{M}}$, denoted by $G(\overline{\mathcal{M}}) = (\overline{\mathcal{V}}, \overline{\mathcal{E}})$. That is, a flow meter measures the transmission line where it is installed and the two buses in both ends; an injection meter measures the bus that is installed, the transmission lines connected to the bus, and all the buses on the other end of the transmission lines. In Fig.

[1] The topology of the power network in Fig. 2 is adapted from the IEEE 14-bus test case system (available online at https://www.ee.washington.edu/research/pstca/, Sept. 2016.)

State estimation protection is closely related to the concept of power network observability. The conventional power network observability analysis studies whether a unique estimate of all unknown state variables can be determined from the measurements.

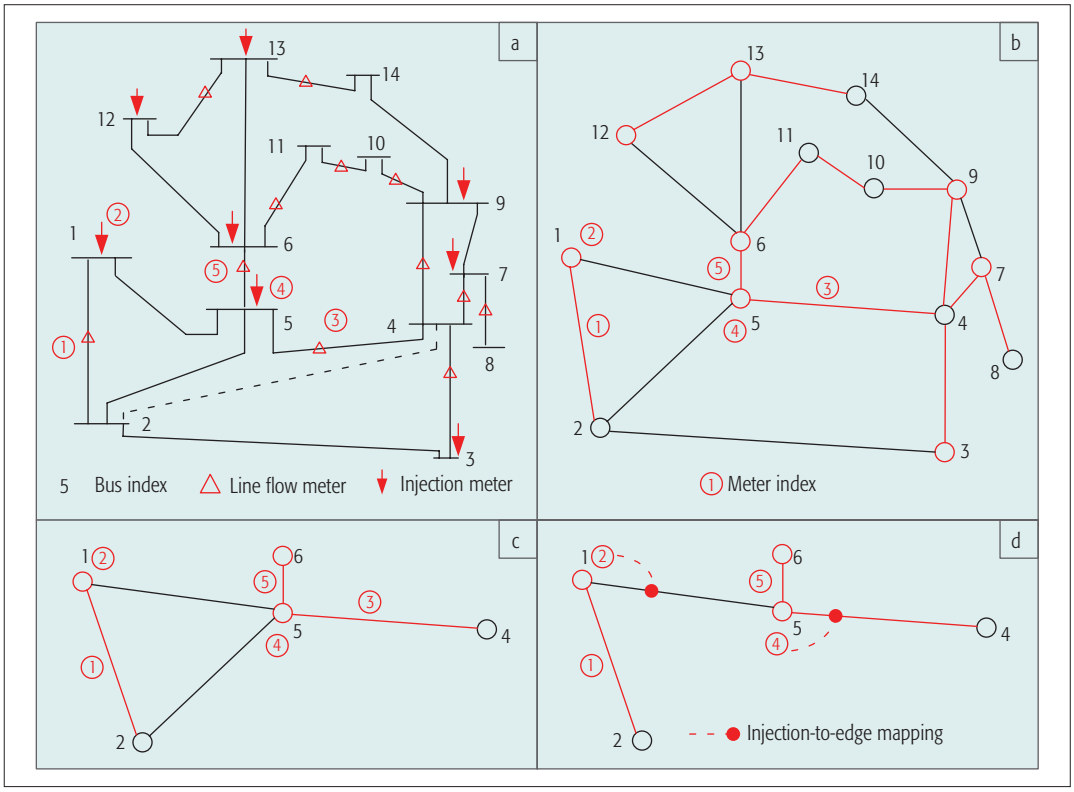


Figure 2. a) An example 14-bus power network and measurements; b) its graph modeling, where the red vertices (edges) denote the buses (transmission lines) that have injection (flow) meters installed; c) an example measured subnetwork; d) an edge-measured Steiner three embedded in the subnetwork.

2c, for instance, the subgraph measured by $\overline{\mathcal{M}} = \{r_1, r_2, r_3, r_4, r_5\}$ is $\overline{\mathcal{V}} = \{\upsilon_1, \upsilon_2, \upsilon_4, \upsilon_5, \upsilon_6\}$ and $\overline{\mathcal{E}} = \{e_{12}, e_{15}, e_{25}, e_{45}, e_{56}\}$, where r_2 and r_4 are injection meters and e_{ij} denotes the edge connecting vertex i and j. For a normal power network, the measured full graph $G(\mathcal{M})$ includes all the vertices $\mathcal{V}$ to estimate all the state variables, but not necessarily all the transmission lines. For instance, we can see that the transmission line between bus 2 and bus 4 is not measured by any meter, and thus is not present in the graph model in Fig. 2b.

Graphical Characterization of State Estimation Protection

State Estimation Problem

The state estimation problem is to derive the unique estimation of θ from the measurements $\mathbf{z}$, which are related by

$$\mathbf{z} = \mathbf{H}\theta + \mathbf{e}.$$

Here, $\mathbf{H}$ denotes the measurement Jacobian matrix and $\mathbf{e}$ denotes independent measurement noise with zero mean. The exact value of $\mathbf{H}$ is related to the physical aspects of the power network, for example, the network topology, the placement of meters, and the transmission line impedance [1]. In particular, we consider in this article a well-functioning power network that a unique estimate $\hat{\theta}$ of the unknown variables can be obtained from the received measurements. This requires a sufficient number of meters to be placed in proper locations such that $\mathbf{H}$ is a full column rank, that is, $rank(\mathbf{H}) = n$. At least n meters are needed to derive a unique state estimation. Meanwhile, the other $m - n$ measurements provide the redundancy to improve the resistance against random errors. Detailed meter placement methods can be found in [13]. Let $\hat{\theta}$ denote the maximum likelihood estimation of θ [1]. The current power systems use a BDD mechanism to remove the bad data, assuming that the errors are random and unstructured. It calculates the residual $\mathbf{r} = \mathbf{z} - \mathbf{H}\hat{\theta}$ and compares its l_2-norm with a prescribed threshold τ. A measurement $\mathbf{z}$ is identified as a bad data measurement if $r = ||\mathbf{z} - \mathbf{H}\hat{\theta}||_2 > \tau$, or otherwise a normal measurement.

Data Injection Attack

A data injection attack compromises the normal measurements through either physical access or remote cyber control, resulting in fabricated measurements $\tilde{\mathbf{z}} = \mathbf{z} + \mathbf{a}$, where $\mathbf{a}$ denotes the injected data. It can be easily shown that an injection attack structured as $\mathbf{a} = \mathbf{H}\mathbf{c}$, where $\mathbf{c}$ is an arbitrary vector, will produce the same BDD residual as the normal measurement z, and thus can introduce a bias c to the state estimate without being recognized as a malicious attack [2]. This kind of attack is commonly referred to as an *undetectable attack*. In general, such an attack requires a high level of coordination to compromise multiple measurements simultaneously. In some cases, however, the adversary can exploit the special structure of $\mathbf{H}$ to achieve the attacking objective by compromising only a small number of measurements. In fact, we will show later how to use graphical methods to exploit the opportunity of an undetectable attack with the minimum number of meter measurements to compromise.

Power Network Observability

State estimation protection is closely related to the concept of power *network observability*. The conventional power network observability analysis studies whether a unique estimate of all unknown state variables can be determined from the measurements [1]. Notice that the observability of a network is related to the network topology and the placement of meter measurements, rather than the value of received measurements in real-time. Out of the m total meters, a set of n meter measurements is referred to as a *basic measurement set* if the estimation of n unknown state variables can be uniquely derived from them. It is proved that the presence of any data injection attack can be detected if we can make sure that the measurements taken from at least one basic measurement set are trustworthy, that is, the meters are well-protected [7]. Intuitively, this is because the estimation obtained from a basic measurement set can be used to validate the result derived from all the meter measurements.

In a large-size power network with several hundred state variables, it could be infeasible to perform a security upgrade to protect n basic measurements with a limited budget. Even with a sufficient budget, protecting the n basic measurements in a random sequence may still open to attackers the possibility to compromise a large number of state variables during the lengthy security installation period. In both cases, it is valuable to devise a method that gives priority to defending a subset of state variables that serve our best interests at the current stage, and offers the possibility of expanding the set of protected state variables in the future.

In light of this, [9] generalizes the concept of power network observability to subnetwork observability. Specifically, a subnetwork $G(\overline{\mathcal{M}}) = (\overline{\mathcal{V}}, \overline{\mathcal{E}})$ is referred to as observable if a unique estimation of $\overline{\mathcal{V}}$ can be derived from $\overline{\mathcal{M}}$. Then, protecting the measurements in $\overline{\mathcal{M}}$ can ensure that any data injection attack can be detected as long as it attempts to compromise any member in $\overline{\mathcal{V}}$. The observability of $G(\overline{\mathcal{M}})$ can be easily determined with a simple matrix calculation. Accordingly, to defend a set of state variables, denoted by $\mathcal{D}$, the problem becomes finding an optimal observable subnetwork $G(\overline{\mathcal{M}}) = (\overline{\mathcal{V}}, \overline{\mathcal{E}})$, either with the minimum number of vertices or the minimum cost to secure the meters in $\overline{\mathcal{M}}$, that satisfies $\mathcal{D} \subseteq \overline{\mathcal{V}}$. An intuitive solution is to enumerate all possible vertices in $\mathcal{S}\backslash\mathcal{D}$ to check if an observable subnetwork can be constructed together with $\mathcal{D}$. This enumeration method, however, is combinatorial in nature, and indeed the problem to find the optimal subnetwork is proved to be NP-Hard [9].

Graphical Characterization of Observability

Alternatively, the network observability has an intuitive characterization using graphs. Specifically, a subnetwork $G(\overline{\mathcal{M}}) = (\overline{\mathcal{V}}, \overline{\mathcal{E}})$ is observable if and only if an *edge-measured Steiner tree* (EMST) [9], denoted by $T = (\overline{\mathcal{V}}, \hat{\mathcal{E}})$, can be constructed from the subnetwork and satisfies the following conditions:

- The reference vertex R is contained in the tree, that is, $R \in \overline{\mathcal{V}}$.
- Each edge $e \in \hat{\mathcal{E}}$ is mapped to a flow meter or an injection meter $p \in \overline{\mathcal{M}}$ that measures it.
- Different edges are mapped to different meters in $\overline{\mathcal{M}}$.

Intuitively, this requires finding a tree that connects all the vertices in the subgraph to the reference vertex, where each edge is mapped to a meter that takes its measurement. For instance, an EMST and the measurement-to-edge mappings are shown in Fig. 2d for the observable subnetwork in Fig. 2c. Such a tree is named a Steiner tree because in general only a subset of vertices is included in the tree. A special case is $\overline{\mathcal{V}} = \mathcal{V}$, where the Steiner tree becomes a *spanning tree* that includes all the vertices in the network [13]. Thanks to the graphical structure of an observable subnetwork, we introduce in the following section some efficient graphical algorithms for security analysis in power systems.

Graph Algorithms for Power System Security Analysis

Maximum-Flow Matching Algorithm

The graphical characterization establishes the equivalence between the subnetwork observability and the existence of an embedded EMST. A natural question is how to construct such an EMST from an observable subnetwork $G(\overline{\mathcal{M}}) = (\overline{\mathcal{V}}, \overline{\mathcal{E}})$, which is very useful in visualizing the network observability to enable efficient tree-based algorithms. As finding a set of meters $\hat{\mathcal{M}} \subseteq \overline{\mathcal{M}}$ to derive a unique estimation of $\overline{\mathcal{V}}$ is easily achievable through a Gauss-Jordan matrix elimination, the question lies in how to find the mappings between $\hat{\mathcal{M}}$ and the edges $\overline{\mathcal{E}}$ to satisfy the EMST definitions. Interestingly, the EMST construction problem can be solved in polynomial time using a maximum-flow method [14].

We use an observable subnetwork in Fig. 2c as an example to illustrate the method to obtain an EMST. As shown in Fig. 2d, we have $\overline{\mathcal{V}} = \{\upsilon_1, \upsilon_2, \upsilon_4, \upsilon_5, \upsilon_6\}$, $\hat{\mathcal{M}} = \{r_1, r_2, r_4, r_5\}$, and the set of edges measured by $\hat{\mathcal{M}}$ is $\overline{\mathcal{E}} = \{e_{12}, e_{15}, e_{25}, e_{45}, e_{56}\}$. Then, a directed graph is constructed in Fig. 3, where υ_1 is chosen as the root to construct the Steiner tree. We select in advance an edge connected to the root, say e_{12}, in the final tree solution. This is achieved by setting both the lower and upper capacity bounds of the edge to be 1. The other edges' lower and upper capacity bounds are set to be 0 and 1, respectively. Then, a maximum flow is calculated from the source(s) to the terminal (t), which is achievable in polynomial time using, for example, the Ford-Fulkerson Algorithm [14]. If the problem is feasible, that is, if the flow solution is 1 in edge e_{12}, we obtain a measurement-to-edge mapping by observing the saturating flows in the graph. Otherwise, the actual EMST solution does not include e_{12} (i.e., the initial guess is wrong), thus we select another edge connected to the root and recalculate the maximum flow problem. Since the subnetwork is observable, the existence of a solution is guaranteed. In the above example, the final measurement-to-edge mapping is $\{r_1, r_2, r_4, r_5\} \leftrightarrow \{e_{12}, e_{15}, e_{45}, e_{56}\}$, while edge e_{25} is not used. Then, the edges obtained by the maximum flow calculation will form a tree that spans all vertices in $\overline{\mathcal{V}}$, as shown in Fig. 2d.

To satisfy the conditions of a feasible EMST, we need to make sure that any selected arc is mapped to a meter that measures it. In particular, if an arc is mapped to an injection meter, all the vertices measured by the injection meter must also be included in the arborescence, as if a pseudo demand is allocated at these vertices.

Commodity Flow Maximization Algorithm

Although finding a minimum EMST that includes a set of vertices $\mathcal{D}$ is NP-Hard, a *commodity flow formulation* that exploits the tree structure of EMST can largely reduce the complexity compared to some enumeration based methods, for example, from several months to a couple of minutes in a medium-size network. Intuitively, this is because the graph-based formulation can significantly reduce the search space of candidate solutions and enable effective off-the-shelf graph/optimization algorithms.

Consider a digraph $\boldsymbol{G}$ =($\mathcal{V}$, $\mathcal{A}$) constructed by replacing each edge in the measured full graph $\overline{G}(\mathcal{M})$ = ($\mathcal{V}$, $\mathcal{E}$) with two arcs in opposite directions. We set the reference bus as the root and allocate one unit of demand to each vertex in $\mathcal{D}$. As shown in Fig. 4, commodities are sent from the root to the vertices in $\mathcal{D}$ through some arcs. Notice that the choice of the vertices $\mathcal{D}$ in Fig. 4 is only for the simplicity of illustration, where an arbitrary subset of vertices $\mathcal{D} \subseteq \mathcal{S}$ can be selected. Then, the vertices in $\mathcal{D}$ are connected to R via the used arcs if and only if all the demands are satisfied. When we require using the minimum number of arcs to deliver the commodity, the used arcs will form a directed tree, referred to as a *Steiner arborescence*. Evidently, the solution to the minimum EMST problem can be obtained if we neglect the orientations of the arcs in the obtained Steiner arborescence. To satisfy the conditions of a feasible EMST, we need to make sure that any selected arc is mapped to a meter that measures it. In particular, if an arc is mapped to an injection meter, all the vertices measured by the injection meter must also be included in the arborescence, as if a pseudo demand is allocated at these vertices. Then, the problem is to satisfy both the actual and pseudo demand using a minimum number of arcs.

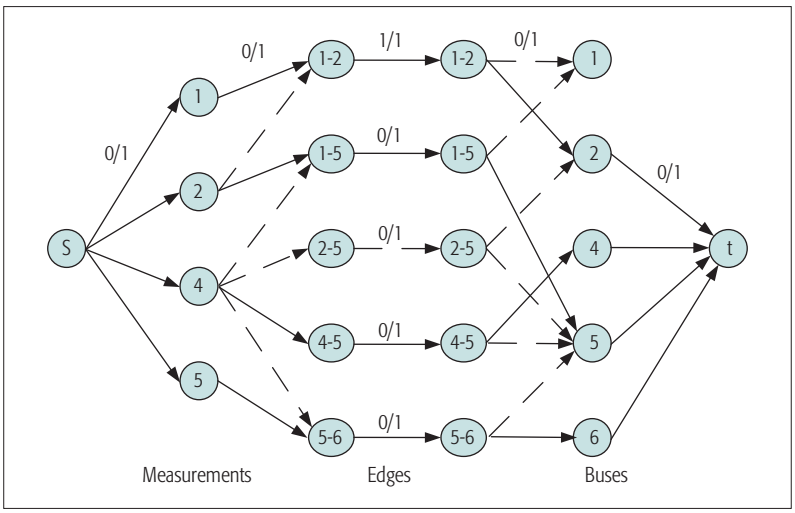


Figure 3. Illustration of maximum flow method for constructing an EMST from an observable subnetwork. The solid lines denote saturating edges while the dashed lines denote unused edges.

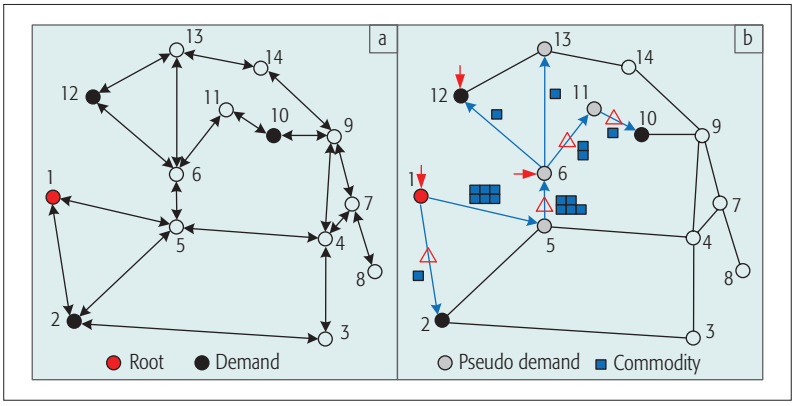


Figure 4. Figure Illustration of commodity flow maximization method for solving the optimal EMST problem: a) the Steiner arborescence constructed from the graph model in Fig. 2b; b) the maximum commodity flow solution when the arc weight is the same for all the arcs.

Based on the commodity flow model, a mixed integer linear programming (MILP) formulation is proposed in [9], and extended to arcs of different weights (different costs are needed to secure the meters) in [12], which can be solved with many off-the-shelf integer optimization tools, such as Gurobi and CPLEX. Accordingly, we can use the mappings from the arcs in the optimal EMST to the optimal set of meter measurements that defends the state variables in $\mathcal{D}$.

Tree Pruning Algorithm

Due to the NP-Hardness of finding an optimal EMST, the commodity flow based method can still result in high computational complexity in a large-size power network consisting of hundreds of buses. A polynomial-time suboptimal algorithm using the idea of tree pruning is considered in [9]. Starting from the full measured graph, the key idea is to iteratively construct an EMST from the subnetwork and prune away redundant vertices not in $\mathcal{D}$, while keeping the remaining subnetwork formed by the residual vertices observable until a shortest possible EMST is obtained. Specifically, the *tree traversal* algorithm can be applied to determine both the sequence and the subset of vertices to be pruned in each iteration.

In Fig. 5, we present an example to illustrate the pruning operation, where a feasible tree containing 12 vertices is presented in Fig. 5a. Vertices 5 and 8 are the *terminal* vertices to be included in the EMST solution. As shown in Fig. 5b, starting from the root υ_1, among the three child vertices of υ_1, only υ_2 can be pruned, since the descendent vertices of either υ_3 or υ_4 contain a terminal vertex. After pruning υ2, we proceed to check υ_3 to see if its child vertex υ_5 can be pruned, which, however, is not feasible because υ_5 is a terminal. Then, we check υ_4, where neither of its child vertices υ_6 and υ_7 can be pruned separately or together. On one hand, this is because υ_6 contains a terminal as its descendent vertices. On the other hand, the removal of υ_7 does not remove the edge e_{46}, which is mapped to the injection meter at υ_6 that measures υ_7, thus resulting in an unobservable residual subnetwork. For υ_7, however, all of its descendent vertices can be pruned as in Fig. 5c. Up to now, we have finished the first round of pruning and obtained a residual tree in Fig. 5d. Then, we use the remaining vertices {υ_1, υ_3, υ_4, υ_5, υ_6, υ_7, υ_8} to generate new EMSTs using the maximum-flow matching algorithm, and repeat the pruning operations iteratively until no vertex can be further pruned.

It is shown in [9] that the tree pruning heuristic (TPH) can achieve comparable performance

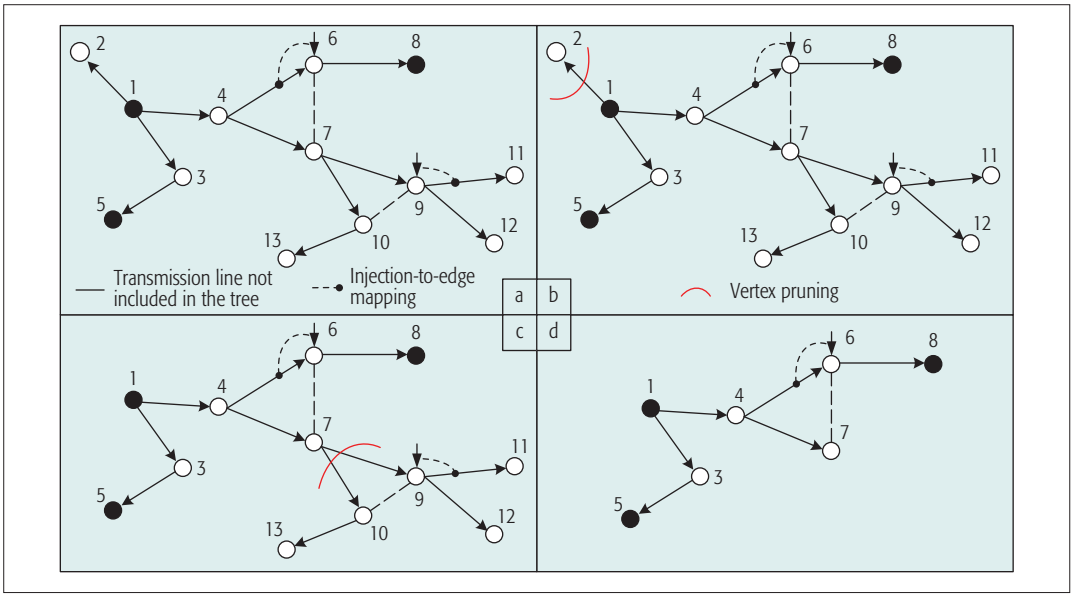


Figure 5. Illustration of the tree pruning algorithm. The shaded vertices are terminals to be included in the EMST.

with the optimal solution obtained from the commodity flow MILP formulation, especially in a large-size network, while inducing much lower complexity. For instance, using a regular computer with Intel Core2 Duo 3.00-GHz CPU and 4 GB of memory, the average computation time needed to solve for $|\mathcal{D}| = 4$ buses out of a 14-bus network is {0.04, 0.2, 0.02} seconds for the arithmetic-based enumeration, the introduced MILP formulation, and the TPH methods, respectively [9]. However, the computation time of arithmetic-based enumeration grows dramatically to around 90 years to solve for $|\mathcal{D}| = 4$ in a 57-bus network, which is computationally infeasible in practice. This, however, takes the MILP and the TPH methods only 3.7 seconds and 0.12 seconds, respectively. As we further increase the size to a 118-bus network, the computation time of the TPH method increase almost linearly to 0.49 seconds, while the optimal MILP formulation increases quickly to around 5 minutes. In this sense, the TPH method can efficiently solve a problem in very large networks of several hundred of buses within a couple of seconds, while it may take the MILP method many days or even months to complete.

Minimum S-T Cut Algorithm

An adversary can also apply graphical methods to exploit the opportunity to launch malicious attacks. A widely used algorithm is the *minimum S-T cut* method, which calculates the minimum sum weights of edges, whose removal would separate a source vertex from a terminal vertex in a weighted graph [10]. Intuitively, an adversary that intends to compromise a state variable will need to separate the corresponding vertex (the terminal) from the reference vertex (the source) in the graph by forming a cut on the edges. Then, the adversary needs to compromise all the meters that measure the edges in the cut. For instance, in Fig. 6a, the cut on e_{78} to attack bus 8 requires the adversary to compromise the flow meter on edge e78 and the injection meter on bus 7. The weight

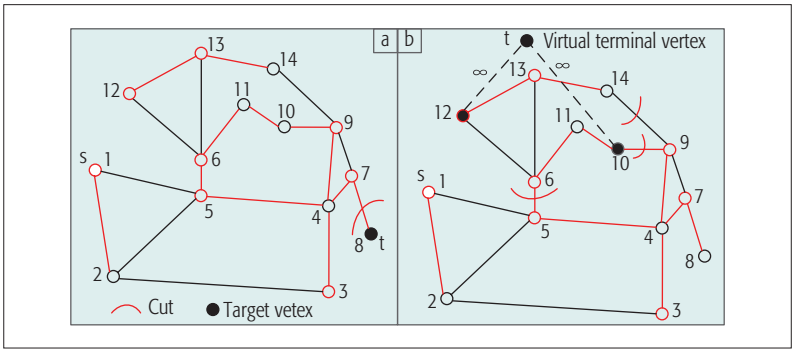


Figure 6. Illustration of minimum S-T cut algorithm to exploit cyber vulnerability. The edge weight is 1 for each edge unless otherwise stated: a) Shows the cut to attack a single bus 8; b) Shows the cut to compromise two buses 10 and 12. In b), a virtual terminal is added to connect the target vertices, while the edge weights between them are set as infinity to avoid a cut on any of them.

of each edge in the calculation of the minimum S-T cut problem can be set as the monetary cost to compromise the meters that measure it.

Similar minimum cut methods can also be applied to compromise a set of state variables [9] (Fig.6b); to find the smallest number of meters that the adversary can control to perform an unobservable attack [4]; to identify the most vulnerable measurements to inject false data [10]; and to exploit the opportunity of data injection attack when some meters are secured or the network topology is only partially known [11, 12]. As the minimum S-T cut can be efficiently calculated in polynomial time, an adversary is able to quickly identify potential network security vulnerability.

Future Research Directions

Application-Oriented Security Analysis

Essentially, the power system state estimation is used for controlling specific applications, such as generation/load power control and electricity price calculation. It is therefore of practical value to perform application-oriented security analy-

As the future smart power grid will integrate a large number of ICT facilities, cyber security is of paramount importance to guarantee the system consistently operating in a secure and efficient state. Graph-based methods are expected to be a set of powerful tools in solving complex cyber security problems in future smart grid.

sis in a higher application layer. Existing studies have shown that data injection attacks that cause blackout and electricity price manipulation have apparent graphical patterns [3, 5]. It is therefore interesting to exploit the underlying graphical structures in the attacks to compromise power applications, such as load prediction, unit commitment, and frequency control. On the other hand, it is also useful to use graphical methods to strategically deploy security countermeasures, for example, to prevent collaborating attacks that compromise the electricity market.

Meter Measurement Placement Optimization

As we are now transforming the legacy power system to the future smart grid, a large amount of electricity infrastructure will be built in the near future, with a mixture of conventional and new metering/communication facilities. Many existing security vulnerabilities often come from the legacy meter measurement placement, which hardly considers the threat of potential collaborating attacks. Graphical methods can be useful to optimize the placement of the meter measurement. By leveraging the graphical properties of network observability, we have the potential to achieve both high state estimation accuracy and high resistance to potential data attacks with relatively low meter placement costs.

Hybrid Graphical and Data-Driven Approaches

Graph-based security analysis is an offline "hardware" approach, where physical protections are performed to ensure the measurements collected from a subset of meters are trustworthy (free from injection attacks). Data-driven attack detection, on the other hand, is an online "software" approach that leverages the statistical features of the measurements/state variables to identify potential abnormal measurements collected from unsecured meters. In particular, The graph-based method is independent of real-time measurements and does not alter the state estimation algorithm in EMS/SCADA. Therefore, it can be potentially combined with data-driven detection to further improve system security. For instance, trustworthy measurements, and hence the subset of trustworthy state estimates derived from them, can be used as side information to improve the detection accuracy of data-driven statistical detections. In general, the graph-based protection method and the data-driven method should be jointly designed.

Security Analysis in the AC Model

Graph algorithms are commonly used to solve linear integer programming problems, and their effectiveness and efficiency to solve security problem in linear DC power systems is unsurprising. In many application scenarios, however, the AC power model, where both voltage amplitude and phase are the state variables, is more preferable than the DC model, for example, to calculate the security constrained optimal power flow. Some studies have shown that data injection attacks to compromise AC state estimation are much more complicated than that in the DC model [15]. On the other hand, the observability of AC state estimation can no longer be characterized as a simple Steiner tree structure as in the DC model. However, network observability may still contain tree-like structures to be identified to defend against potential attacks on AC state estimation.

Conclusions

In this article, we have provided a graphical framework for performing security analysis in power system state estimation. From the perspective of both the system operator and the adversary, we have introduced several effective graph-based algorithms to solve security problems in state estimation. Compared to the commonly used arithmetic-based security analysis, graph-based analysis helps visualize some complex problem structures, which can lead to efficient optimal or reduced-complexity suboptimal graph-based algorithms. As the future smart power grid will integrate a large number of ICT facilities, cyber security is of paramount importance to guarantee that the system is consistently operating in a secure and efficient state. Graph-based methods are expected to be a set of powerful tools in solving complex cyber security problems in the future smart grid.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Project number 61501303) and the Foundation of Shenzhen City (Project number JCYJ20160307153818306). The work of Y. J. Zhang is supported in part by General Research Funding (Project number 14200315) from the Research Grants Council of Hong Kong and the Theme-Based Research Scheme (Project number T23-407/13-N).

References

[1] A. Abur and A. G. Expósito, *Power System State Estimation: Theory and Implementation*, New York: Marcel Dekker, 2004.
[2] Y. Liu, P. Ning, and M. Reiter, "False Data Injection Attacks Against State Estimation in Electric Power Grids," *Proc. ACM CCS*, Chicago, USA, Oct. 2009.
[3] S. Bi and Y. J. Zhang, "False Data Injection Attack to Control Real-Time Price in Electricity Market," *Proc. IEEE Globecom*, Atlanta, USA, Dec. 2013.
[4] O. Kosut *et al.*, "Malicious Data Attack on the Smart Grid," *IEEE Trans. Smart Grid*, vol. 2, no. 4, Dec. 2011, pp. 645–58.
[5] Y. Yuan, Z. Li, and K. Ren, "Modeling Load Redistribution Attacks in Power Systems," *IEEE Trans. Smart Grid*, vol. 2, no. 2, June 2011, pp. 382–90.
[6] Y. Huang *et al.*, "Bad Data Injection in Smart Grid: Attack and Defense Mechanisms," *IEEE Commun. Mag.*, vol. 51, no. 1, Jan. 2013, pp. 27–33.
[7] R. Bobba *et al.*, "Detecting False Data Injection Attacks on Dc State Estimation," *Proc. CPSWEEK*, Apr. 2010.
[8] S. Bi and Y. J. Zhang, "Defending Mechanisms Against False-Data Injection Attacks in the Power System State Estimation," *Proc. IEEE Globecom*, Houston, USA, Dec. 2011.
[9] S. Bi and Y. J. Zhang, "Graphical Methods for Defense Against False-Data Injection Attacks on the Power System State Estimation," *IEEE Trans. Smart Grid*, vol. 5, no. 3, May 2014, pp. 1216–27.
[10] K. C. Sou, H. Sandberg, and K. H. Johansson, "Computing Critical K-Tuples in Power Networks," *IEEE Trans. Power Syst.*, vol. 27, no. 3, Aug. 2012, pp. 1511–20.
[11] M. A. Rahman and H. Mohsenian-Rad, "False Data Injection Attacks with Incomplete Information Against Smart Power Grids," *Proc. IEEE Globecom*, Anaheim, USA, Dec. 2012.
[12] S. Bi and Y. J. Zhang, "Using Covert Topological Information for Defense Against Malicious Attacks on DC State Estimation," *IEEE JSAC*, vol. 32, no. 7, July 2014, pp. 1–15.
[13] G. R. Krumpholz, K. A. Clements, and P. W. Davis, "Power System Observability: A Practical Algorithm Using Network Topology," *IEEE Trans. Power App. Syst.*, vol. PAS-99, no. 4, Jul. 1980, pp. 1534–42.

[14] A. Barglela, M. R. Irving, and M. J. H. Sterling, "Observability Determination in Power System State Estimation Using a Network Flow Technique," *IEEE Trans. Power Syst.*, vol. PWRS-1, no. 2, May 1986.
[15] G. Hug and J. A. Giampapa, "Vulnerability Assessment of AC State Estimation with Respect to False Data Injection Cyber-Attacks," *IEEE Trans. Smart Grid*, vol. 3, no. 3, Sept. 2012, pp. 1362–70.

Biographies

Suzhi Bi [S'10, M'14] (bsz@szu.edu.cn) received the B.Eng. degree in communications engineering from Zhejiang University, Hangzhou, China, in 2009, and the Ph.D. degree in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. From 2013 to 2015, he was a research fellow in the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He is now an assistant professor with the College of Information Engineering, Shenzhen University, Shenzhen, China. His current research interests include wireless information and power transfer, medium access control in wireless networks, and smart power grid communications.

Angela Yingjun Zhang [S'00, M'05, SM'11] (yjzhang@ie.cuhk.edu.hk) received her Ph.D. degree in electrical and electronic engineering from the Hong Kong University of Science and Technology in 2004. Since 2005, she has been with the Department of Information Engineering, Chinese University of Hong Kong, where she is currently an associate professor. She was with the Wireless Communications and Network Science Laboratory at the Massachusetts Institute of Technology during the summers of 2007 and 2009, and with Netlab at the California Institute of Technology during the summer of 2014. Her research interests mainly include wireless communications systems and smart power systems, in particular optimization techniques for such systems. She is an executive editor of *IEEE Transactions on Wireless Communications*, and an associate editor of *IEEE Transactions on Communications*. Previously, she served many years as an associate editor of *IEEE Transactions on Wireless Communications* and *Security and Communications Networks* (Wiley), and as a Guest Editor of a Feature Topic in *IEEE Communications Magazine*. She has served as a Workshop Chair of IEEE ICCC 2014 and 2013, TPC Vice Chair of the Wireless Networks and Security Track of IEEE VTC 2014, TPC Vice-Chair of the Wireless Communications Track of IEEE CCNC 2013, TPC Co-Chair of the Wireless Communications Symposium at IEEE GLOBECOM 2012, Publication Chair of IEEE TTM 2011, TPC Co-Chair of the Communication Theory Symposium at IEEE ICC 2009, Track Chair of ICCCN 2007, and Publicity Chair of IEEE MASS 2007. She was a Co-Chair of the IEEE ComSoc Multimedia Communications Technical Committee and the IEEE Communication Society GOLD Coordinator. She was a co-recipient of the 2014 IEEE ComSoc APB Outstanding Paper Award, the 2013 IEEE SmartgridComm Best Paper Award, and the 2011 IEEE Marconi Prize Paper Award on Wireless Communications. She was the recipient of the Young Researcher Award from the Chinese University of Hong Kong in 2011. As the only winner from engineering science, she won the Hong Kong Young Scientist Award 2006, conferred by the Hong Kong Institution of Science. She is a Fellow of IET.

ACCEPTED FROM OPEN CALL

Full-Duplex Cellular Networks

Rongpeng Li, Yan Chen, Geoffrey Ye Li, and Guangyi Liu

Before putting FD networking into practice, we need to understand to which scenarios FD communications should be applied under the current technology maturity, how bad the performance will be if we do nothing to deal with the newly introduced interference, and most importantly, how much improvement could be achieved after applying advanced interference management solutions.

ABSTRACT

Full-duplex (FD) communications with simultaneous transmission and reception on the same carrier have long been deemed a promising way to boost spectrum efficiency, but hindered by the techniques for self-interference cancellation (SIC). Recent breakthroughs in analog and digital signal processing yield the feasibility of over 100 dB SIC capability, and make it possible for FD communications to demonstrate nearly doubled spectrum efficiency for point-to-point links. Now it is time to shift at least partially our focus to FD networking, such as in cellular networks. FD networking has more complicated interference environments. Therefore, its performance improvement is not that straightforward compared with half-duplex networking. Before putting FD networking into practice, we need to understand to which scenarios FD communications should be applied under the current technology maturity, how bad the performance will be if we do nothing to deal with the newly introduced interference, and most importantly, how much improvement could be achieved after applying advanced interference management solutions. We will discuss all these questions in this article. In particular, we will investigate advanced interference management solutions, including power control and user scheduling, and show that up to 91 percent spectrum efficiency gain and 110 percent energy efficiency gain of FD cellular networks over its HD counterpart can be achieved by applying these solutions.

INTRODUCTION

To satisfy the surging traffic demand, mobile networks are facing unprecedented challenges to further improve their efficiency of spectrum usage. Currently, mobile networks operate in a half-duplex (HD) mode, which implies only one direction transmission on a frequency carrier at any time and no extra cost for spatial separation. For example, the base station (BS) can transmit to users (downlink (DL)) at one time and frequency radio resource, and receive from users (uplink (UL)) at another. These time and frequency radio resources are also known as channels. They can be separated by time or frequency dimension, called time-division duplex (TDD) or frequency-division duplex (FDD) mode, respectively. On the other hand, transmission and reception on the same channel at the same time, also known as full-duplex (FD) communication, has long been dreamed of but has been hindered by strong self-interference from a node's transmitter to its receiver. In an FD transceiver, the self-interfering signal from its transmitter is usually 100 dB stronger than the intended receiving signal. As hard as trying to hear a whisper while shouting at the top of your lungs, strong self-interference in an FD system will easily cause the radio chain at the receiver to be saturated [1] and unable to work properly, not to mention decoding the data.

However, recent breakthroughs in analog and digital signal processing facilitate the real application of FD communications. It is now feasible to have up to 110 dB self-interference cancellation (SIC) capability [2]. Therefore, self-interference is mostly removed with the residual strength reduced to the same level as the signal of interest before going through the decoding chain at the receiver, which makes data decoding feasible. As a result, there have been many real-time FD prototypes reported [2–5].

While roughly doubled throughput has been reported for single-link FD transmission [5], the performance improvement of FD networks is not that straightforward due to the new interference introduced by FD links. The deployment of FD networking needs to consider the following two factors. First, it is still costly to equip FD functionality with above 100 dB for all user equipment (UE), so most of the UEs may still work in HD mode at least in the near future. Therefore, we assume only BSs work in FD mode. Second, coexistence of both UL and DL transmission on the same channel at the same time in all cells introduces far more complicated interference, as illustrated in Fig. 1. From Fig. 1, besides the inter-cell BS-to-UE and UE-to-BS interference that already exists in HD networks, dynamic TDD networks[1] [6] and FD networks experience extra inter-cell inter-BS and inter-UE interference. Furthermore, FD networks face intra-cell inter-UE interference as well as residual self-interference after SIC. Hence, smart interference management techniques are necessary to deal with various types of interference and ensure performance improvement of FD networks compared with HD networks [7–11]. Therefore, given the current SIC and interference management capability, it is critical to carefully select application scenarios for FD communications and design protocols and algorithms to deal with the newly introduced interference.

SCOPE AND KEY FINDINGS

In this section, we will first introduce the metrics for the performance evaluation of FD cellular networks and then briefly summarize our findings to facilitate FD cellular networks.

[1] In dynamic TDD networks, different BSs are designed to have the flexibility to configure different UL and DL subframe patterns, which aims to better adapt to dynamic variations in DL/UL traffic demands on the BS basis.

Digital Object Identifier:
10.1109/MCOM.2017.1600361CM

Rongpeng Li is with Zhejiang University, and was with Huawei Technologies Co., Ltd.; Yan Chen is with Huawei Technologies Co.; Geoffrey Ye Li is with Georgia Institute of Technology; Guangyi Liu is with China Mobile Research Institute.

Scope and Evaluation Framework

There is a wide consensus that applying FD communications to macrocells is not a good candidate scenario because of the large transmission power of macro BSs imposed by the large coverage requirement [9]. Direct calculation yields that above 140 dB SIC is required to bring down the transmission signal to a level of –100 dB [9] for the macro BSs transmitting at 46 dBm. Instead, the architectural progression toward short-range systems, such as small-cell (e.g., picocells) systems where the cell-edge path loss is less than that in macrocell systems, makes the self-interference reduction problem much more manageable. Therefore, in this article, we focus on cellular networks with pico BSs operating in FD mode while leaving macro BSs and UEs in HD mode. We will analyze how serious the problem could be if we directly introduce FD communications to the pico BSs in heterogeneous networks, and how effective different interference management strategies may be.

We use the two important indicators for system performance evaluation, i.e., system spectrum efficiency (SE) and system energy efficiency (EE). The system SE is defined as the overall UL and DL throughput per unit bandwidth. Mathematically, it is given by

$$\mathrm{SE} = \frac{T_{\mathrm{tot}}^{\mathrm{UL}} + T_{\mathrm{tot}}^{\mathrm{DL}}}{B_{\mathrm{tot}}}, \quad (1)$$

where $T_{\mathrm{tot}}^{\mathrm{UL}}$, T_{tot} and $B_{\mathrm{tot}}^{\mathrm{DL}}$ indicate the UL and DL throughput and the allocated bandwidth, respectively. On the other hand, the system EE is defined as the aggregated bits transmitted in both UL and DL in unit bandwidth per joule energy consumed. Here we only consider transmission energy and ignore signal processing energy since the previous "air interface" radiated power is more tightly related to interference management strategies involved in the article. Then it could be mathematically formulated as

$$\mathrm{EE} = \frac{\left(T_{\mathrm{tot}}^{\mathrm{UL}} + T_{\mathrm{tot}}^{\mathrm{DL}}\right) \times T_i / B_{\mathrm{tot}}}{E_{\mathrm{tot}}^{\mathrm{UL}} + E_{\mathrm{tot}}^{\mathrm{DL}}} = \frac{SE}{P_{\mathrm{tot}}^{\mathrm{UL}} + P_{\mathrm{tot}}^{\mathrm{DL}}}, \quad (2)$$

where E_{tot} and P_{tot} stand for energy and power consumption, respectively, T_i denotes the transmission time, thus $E_{\mathrm{tot}} = P_{\mathrm{tot}} \times T_i$. In Eq. 2, the superscripts UL and DL are used to indicate the UL and DL energy, power, and throughput. From Eq. 2, system SE and EE are correlated. It has been demonstrated that there exists an interesting SE-EE tradeoff relationship in different types of networks, since the maximization of total SE and the minimization of the total P_{tot} are usually not achieved at the same time. In this article, we will investigate the behavior of such a relationship in FD networks.

Key Findings

In the rest of the article, we start our evaluation from a single-cell FD network. An optimization problem is formulated to maximize the system SE with the transmission power and user selection as control variables. We will show a surprising observation from the analytical solution, that for a given pair of UL and DL UEs, the power control for both the BS and the selected UL UE has a binary feature, i.e., either transmitting at its full power level or completely muting. Based on this observation, a joint power control and user selection problem reduces to a UE scheduling problem only, and interference awareness will play an important role in such a process. As a step further, we investigate multi-cell FD networks and identify the dominant interference for different network configurations based on system level simulations. We will demonstrate through system SE and EE evaluation that up to 91 percent SE gain and 110 percent EE gain can be achieved with different interference management schemes. In other words, FD networking will work, at least for the considered heterogeneous network setting with 110 dB SIC capability.

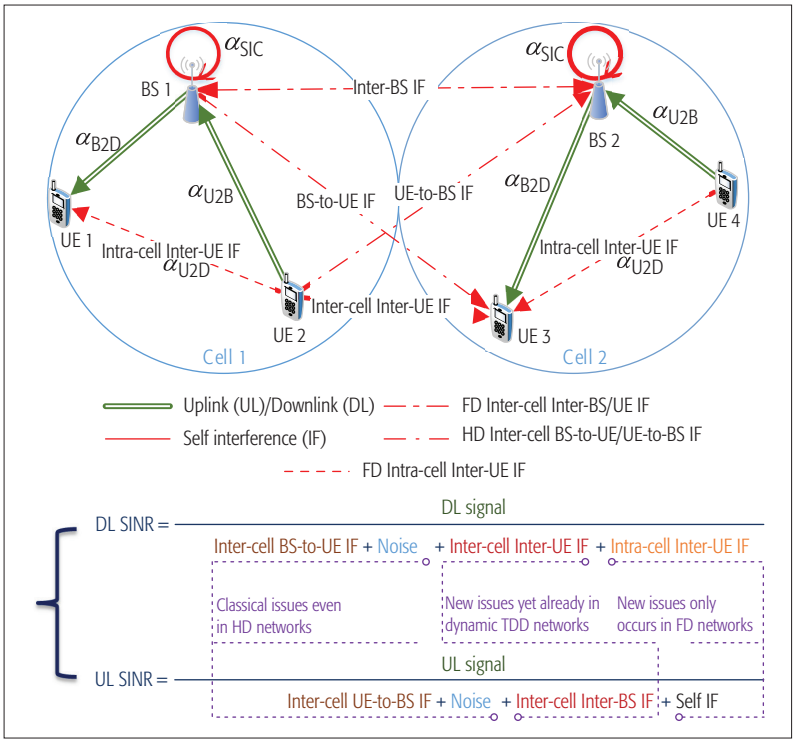


Figure 1. An illustration of different types of interference in FD networking.

Single-Cell FD Network

In a single-cell FD network, the interference situation is much less complicated. However, as shown in Fig. 1, the FD network still needs to deal with the intra-cell interference from UL UE to DL UE and the residual self-interference at the transceiver of the BS. In this case, the problem to maximize the total system throughput of both UL and DL can be formulated as follows,

$$\max_{i,j,P_i^{\mathrm{DL}},P_j^{\mathrm{UL}}} \quad f = \log\left(1 + \frac{\alpha_{\mathrm{B2D}} P_i^{\mathrm{DL}}}{N_0 + \alpha_{\mathrm{U2D}} P_j^{\mathrm{UL}}}\right) + \log\left(1 + \frac{\alpha_{\mathrm{U2B}} P_j^{\mathrm{UL}}}{N_0 + \alpha_{\mathrm{SIC}} P_i^{\mathrm{DL}}}\right)$$

$$s.t. \qquad 0 < P_i^{\mathrm{DL}} \le P_{\max}^{\mathrm{DL}}, 0 < P_j^{\mathrm{UL}} \le P_{\max}^{\mathrm{UL}}, \quad (3)$$

where N_0 denotes the noise power, P_i^{DL} and P_j^{UL} denote the transmission power of the BS (to DL UE i) and UL UE j, and are limited by the corresponding maximum values $P_{\max}^{\mathrm{DL}}$ and $P_{\max}^{\mathrm{UL}}$, respectively, α_{B2D}, α_{U2D}, and α_{U2B} characterize the

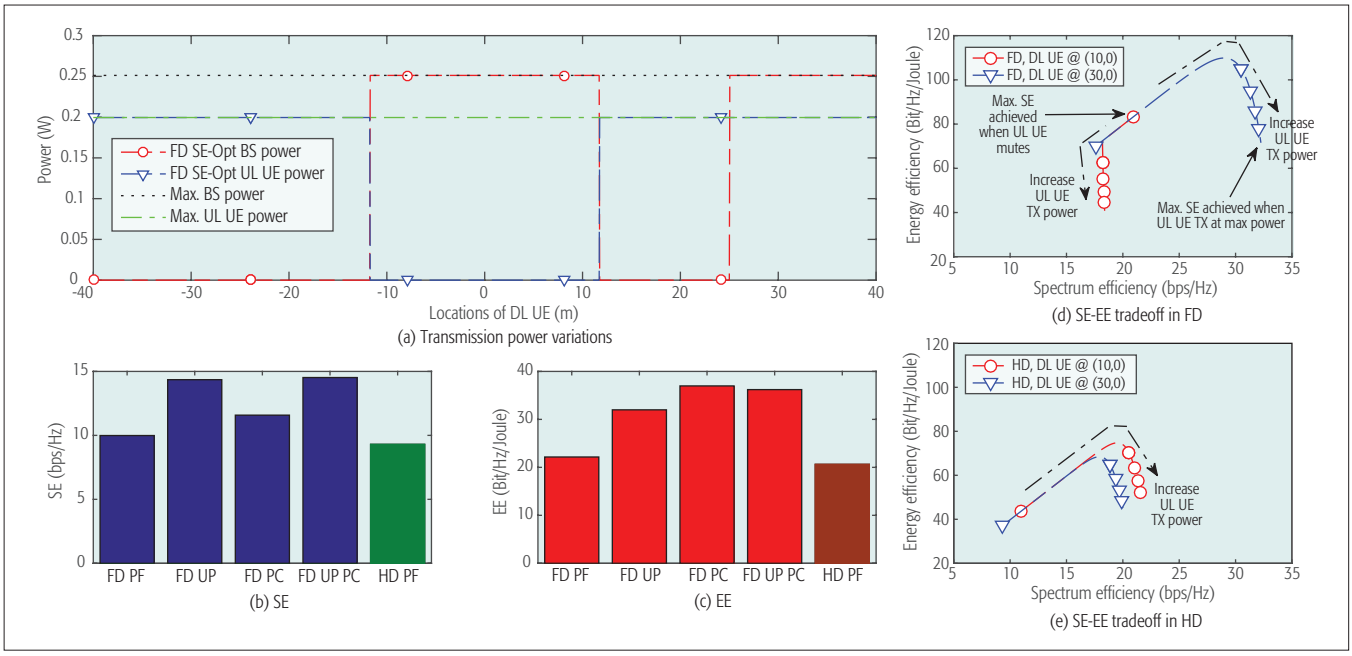


Figure 2. Performance of single-cell network: a) the optimal transmission power of the pico BS and the UL UE in terms of SE maximization; b, c) the SE and EE performance improvement of FD network over HD network under power control (PC) and/or user pairing (UP); d, e) the SE-EE relation for with a given pair of UL and DL UEs at different positions.

Key Findings: (1) For a given pair of UEs, the SE-optimized power control result has a binary feature. (2) The SE-EE relationship in single-cell FD network is dependent on the positions of UEs and might be different from that in HD network.

channel power gains from the BS to the DL UE, from the UL UE to the DL UE, and from the UL UE to the BS, respectively, and will be affected by UL and DL UE scheduling. α_{SIC} indicates the SIC capability. Notably, we do not consider fast fading in these channels, and also regard the self-interfering channel after SIC as a line-of-sight channel with the pathloss equaling the SIC capability.[2]

In this section, we will first dive into the power control problem with only one given pair of UL and DL UEs, and then extend to a multiple UE situation and investigate the problem of UE scheduling. Finally, we will provide the system-level analyses for its SE and EE performance, as well as the tradeoff between them.

Binary Power Control

For a given UL and DL UE pair, the problem in Eq. 3 reduces to joint optimization of the transmission powers of the BS and the UL UE. Without loss of generality, when the transmission power of the UL UE, P_j^{UL}, is fixed, by taking the derivative of Eq. 3 with respect to the transmission power of the BS, P_i^{DL}, we can find that there exists at most one minimum point and no maximum point in the interval $[0, P_{max}^{DL}]$ for the function in Eq. 3. Therefore, the optimal value for P_i^{DL} to maximize the sum rate lies at the two end points of the interval, i.e., 0 or P_{max}^{DL}. Hence, the BS either transmits no signal to turn the network into HD mode or with the maximum power level. A similar result can be obtained for the transmission power of the UL UE. The joint optimization of both, as a generalized case, offers three solution candidate pairs: $(0, P_{UEmax})$, $(P_{BSmax}, 0)$, and (P_{BSmax}, P_{UEmax}), which shows exactly the binary feature and demonstrates appealing computational efficiency to obtain the solution in Eq. 3.

To demonstrate the binary power control feature more clearly, Fig. 2 depicts our simulation results. In our simulation, the pico BS and the UL UE are located at (0,0) and (–25,0), respectively, and other parameters are set as in Table 1. By moving the DL UE along the horizontal axis from (–40,0) to (40,0), we show the optimal power control solutions for both the UL UE and the BS under 110 dB SIC capability [2] in Fig. 2a. Here, instead of applying our analytical observation above, we perform optimization by exhaustive search. From Fig. 2a, to achieve the maximum system SE, the BS and the UL UE either transmit at their maximum power levels or just mute to fall back to HD mode, which is consistent with our analytical observation. Moreover, Fig. 2a also implies that along with the moving of the DL UE, the system will fall back to HD mode for most of the DL UE positions. Hence, for the network with multiple UEs, it is essential to schedule one UL UE and one DL UE to form a pair in FD mode and thus obtain a larger SE gain.

Interference-Aware User Scheduling

Given the binary feature of power control, the SE maximization problem in Eq. 3 reduces to a UE scheduling problem, namely, for a given time-frequency resource, how to select one UL UE and one DL UE from all active UEs to properly work together. Basically, there have been many existing scheduling methods in the HD network, such as proportional fairness (PF) [7] and round-robin, which the FD network could directly take advantage of. For example, the FD network could follow the standard PF procedure to select DL UE and UL UE independently and pair them. However, the ignorance of inter-UE interference in such a method could degrade the performance. Further-

[2] Usually, Rician fading has been leveraged to model the channel after analog cancellation [12], and it is also found that the residual self-interference after SIC follows circular symmetric complex normal distribution.

more, our simulation results will show that interference awareness should be an important feature for the UE pairing process. There are different levels of interference awareness and also various procedures to achieve that awareness. If we can track the inter-user interference channel fast enough, the short-term interference can be captured, which would be best for performance but with the highest overhead in tracking such information. On the other hand, we may only exploit long-term statistics of interference, such as the path-loss, which can be easily derived from the relative user positions. In this case, the interference-aware user scheduling problem turns into a distance-aware problem, and has been investigated in [11, 13].

Here we give an example of a distance-aware joint PF UE pairing algorithm, in which the BS takes turns to select the first user sorted by the PF criterion in the UL or the DL, and then pairs a DL or UL UE with the largest distance. To show the benefit of the binary power control (PC for short) and distance-aware joint PF UE pairing (UP for short) schemes, we simulate a single-cell FD network with eight randomly deployed UL or DL UEs. Without loss of generality, the baseline HD network in our simulation is assumed to work in FDD mode, i.e., the UL or the DL uses half of the total bandwidth. Meanwhile, despite the existence of advanced user scheduling methods applicable to the HD networks as well, we only consider the standard PF method for the HD networks here as an example. The system SE and EE under different strategies are shown in Fig. 2b and Fig. 2c, respectively. From the figures, the FD network shows trivial gain over the traditional HD network without power control or UE pairing. However, when either power control or UE pairing is used, the performance gain can be significantly improved. In particular, the joint power control and UE pairing scheme can provide around 45 percent and 60 percent boost in system SE and EE, respectively.

SE-EE Relationship

As before, we consider one BS with a given UL and DL UE pair but at different locations and find the SE-EE relationship by varying the transmission power of the UL UE from 0 to 23 dBm while fixing the transmission power of the BS. Figure 2d and Fig. 2e demonstrate the SE-EE relationship for the FD and HD networks, respectively. From the figures, the shape of the SE-EE relation does not change with the UE locations in the HD network since there is no interference between the UL UE and the DL UE. However, due to the inter-user interference in the FD network, the relative location between the UL and DL UEs significantly affects the SE-EE relationship, which confirms the effectiveness of the proposed UE pairing method with location awareness. Moreover, for different UE locations, the maximum system SE is derived either when the UL UE transmits at its maximum power or when it is completely muted, which again aligns with our binary power control results in Fig. 2a. In brief, different from the HD network, the SE-EE relationship for the FD network will be dependent on the positions of the UEs. Nevertheless, with advanced interference management strategies in the FD network, EE performance can be improved by around 60 percent when the maximized SE is increased by around 45 percent.

Category	Sub-category	Configuration
TTI		1 ms
Bandwidth		UL or DL in HD: 10 MHz; UL or DL in FD: 20 MHz
Topology	Macro	500 m inter-site distance (ISD) at static positions with three sectors
	Pico	3, 6, …, or 18 picos uniformly distributed in 500 m-ISD macro's region
	UE	*Uniform:* 192 users uniformly distributed in 500 m-ISD macro's region
		Clustered: eight users uniformly distributed in 40 m-radius picocell's region
Propagation model	Pathloss	Strictly following Table A. 1-3 in 3GPP TR 36.828 [14]
	Shadowing	Macro to pico: 6 dB; macro to UE: 10 dB; pico to UE: 10 dB
		UE to UE: 12 dB; pico to pico: 6 dB
	Noise figure	Macro: 5 dB; pico: 13 dB; UE: 9 dB
Maximum transmission power		Macro: 46 dBm; pico: 24 dBm; UE: 23 dBm
SIC capability		50 dB to 120 dB, 110 dB by default
Cell range extension (bias)		6 dB
Proportional fairness		Window length: 500; exponent factor: 0.05

Table 1. Main parameters in the system-level simulator, which are compatible with 3GPP TR 36.828 [14].

Multi-Cell FD Networks

In this section, we investigate multi-cell FD networks. As illustrated in Fig. 1, multi-cell FD networks suffer from more complicated interference. Therefore, we first take a look at how bad the interference situation is and which type of interference is dominant. Then we will discuss which solution is most effective, especially to deal with the dominant interference, and how much gain we might expect in terms of system SE and EE from the FD networks.

As mentioned earlier, we consider a multi-cell heterogeneous network with the macro BSs working in HD mode and the pico BSs working in FD mode. System-level simulations are used to answer the aforementioned questions. Specifically, seven macro BSs in total are located at the vertices and the center of a hexagon, and the pico BSs are randomly scattered in each sector of the macro BSs [14]. The system parameters are listed in Table 1. Moreover, we consider two network configurations:

- *Uniform Case*: UEs are uniformly dropped in the coverage of the macro BSs and associate with the macro BSs or the pico BSs following the standard strongest received signal strength (RSS) criterion. Besides, cell range expansion toward the pico BSs is leveraged by virtually adding 6 dB bias to the received power of the pico BSs. Moreover, the macro BSs and the pico BSs operate in the same band.
- *Clustered Case*: UEs are uniformly dropped in the coverage of the pico BSs and only associated with the pico BSs. In other words, randomly distributed UEs form different clusters, and the positions of UEs in each cluster are limited to the range of one pico BS. The macro BSs and the pico BSs operate in different bands.

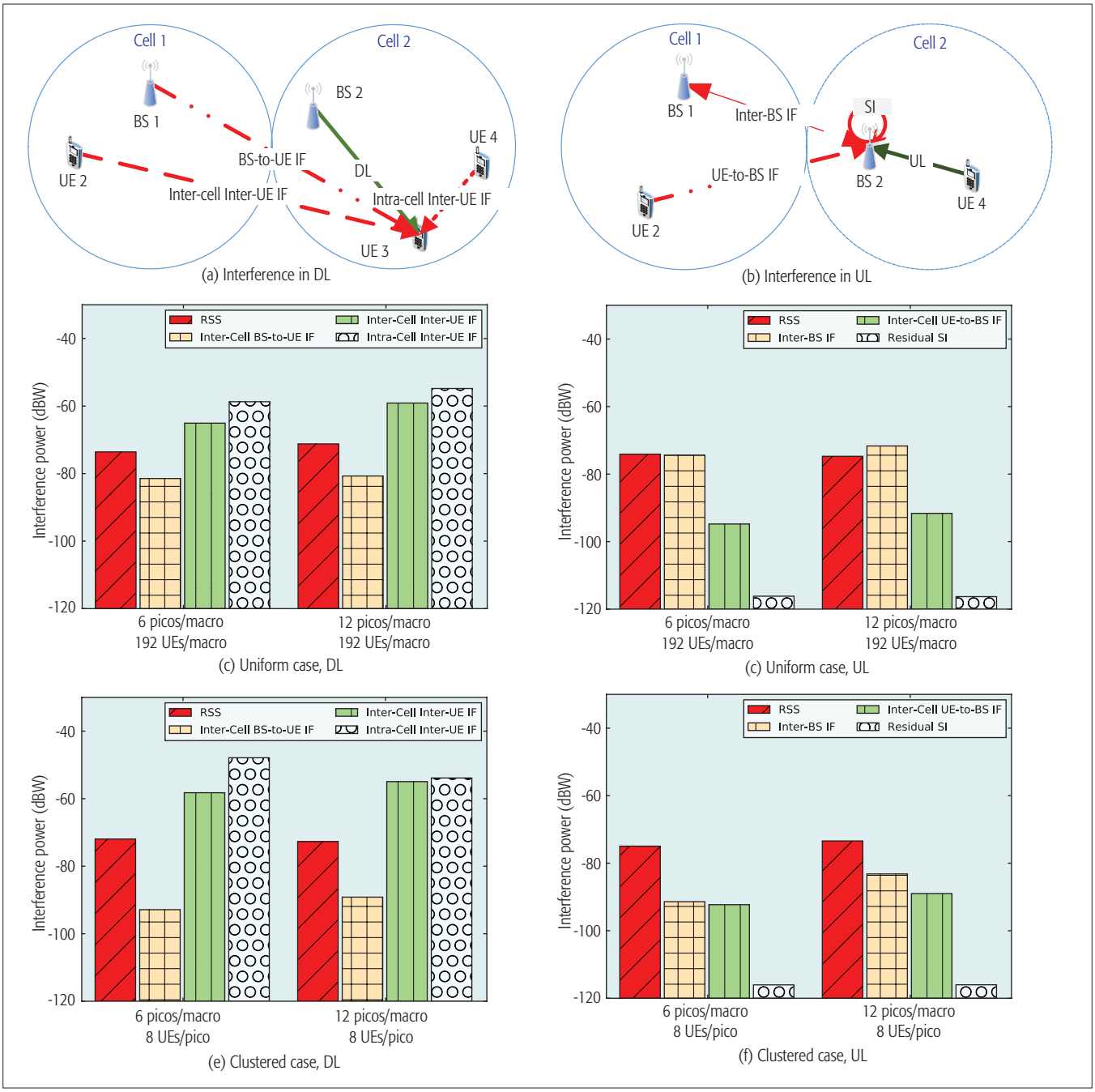


Figure 3. a) Interference in DL; b) UL and corresponding powers in two typical scenarios: c, d) uniform case; e, f) clustered case. Key findings: The intra-cell inter-UE interference dominates in DL, while UL transmission suffers in cochannel heterogeneous networks.

As before, the results are averaged over 100 user drops for both cases.

Interference Analyses

In this section, we investigate the strength of different types of interference for both network configurations by exploiting the standard PF scheduling method for UL and DL UEs separately and not applying any smart interference management scheme. Furthermore, since the interference situation is different, we present the results separately in Fig. 3a and Fig. 3b. We shall see which direction is affected more seriously and which interference is more dominant.

Uniform Case: Figure 3c and Fig. 3d show the interference powers for the UL and DL UEs, respectively. Two groups of results are shown in each figure, corresponding to the two settings (i.e., 6 and 12) of the pico BS density for each macrocell.[3]

From Fig. 3c, for DL transmission, the strongest interference in most cases is from the UL UE of the same cell, which is a unique problem in the FD networks. Meanwhile, DL transmission, on average, is affected more by inter-cell inter-UE interference than by inter-cell BS-to-UE interference for the first group.[4] On the other hand, for UL transmission, the interference from the neighboring BSs, including both the pico BSs and the macro BSs, dominates, as demonstrated by the first group of results in Fig. 3d. Inter-cell interference power increases with the number of pico

[3] As may be needed for result comparison, when there are 192 uniformly distributed UEs per macro BS, statistically around four to eight UEs are associated to each pico BS.

[4] Given the random drop of UEs, the second-order statistic also shows that the inter-cell inter-UE interference has much larger dynamic range that that of inter-cell BS-to-UE interference.

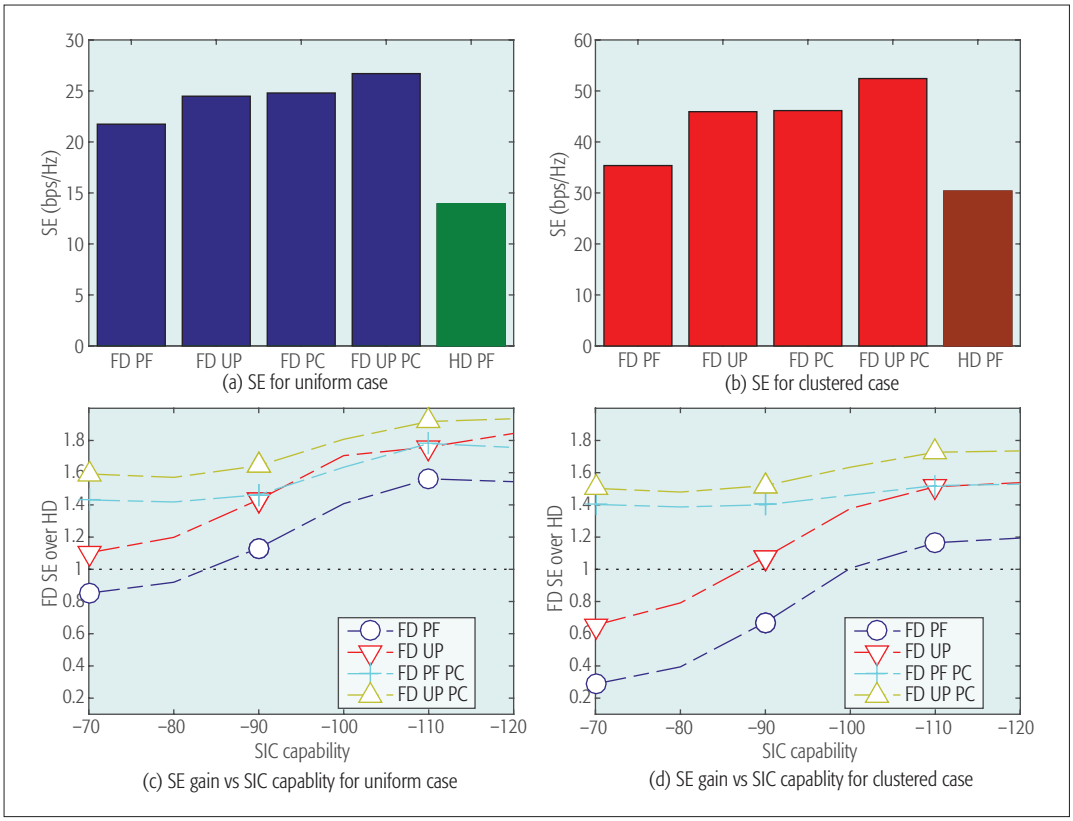


Figure 4. SE performance between uniform case (6 pico BSs/Macro BS, 192 UEs/Macro BS) and clustered case (6 pico BSs/Macro BS, 8 UEs/Pico BS): a, b) the SE under power control (PC) and/or user pairing schemes, 110 dB SIC assumed; c, d): the SE gain of FD networks over HD networks versus the SIC capability.

Key findings: (1) The combination of power control and user pairing algorithm effectively mitigates the annoying interference and provides 91% and 72% SE gain for the uniform and clustered case, respectively; (2) The minimum required SIC capability to obtain SE gain in FD networks are 83 dB and 100 dB for the uniform and clustered case, respectively.

BSs. Moreover, it also implies that more users in the FD networks will incur larger inter-cell inter-UE interference for both UL and DL from Fig. 3c and Fig. 3d.

Clustered Case: In this scenario, we discuss the impact of interference when UEs are clustered. The corresponding results are shown in Fig. 3e and Fig. 3f. Compared with the uniform case, both the inter-cell BS-to-UE interference in the DL and the inter-BS interference in the UL becomes significantly smaller due to the absence of the macro BSs. However, the inter-UE interference is still very strong and needs interference management schemes, so as to exploit the potential benefit of FD communications. Moreover, similar to that in the uniform distribution case, along with the increase in the number of the pico BSs, the interference problem becomes more severe.

Network SE and EE

In this section, we will investigate how the interference management schemes in the single-cell FD network could contribute to improving the multi-cell performance in terms of SE and EE and provide the corresponding results in Fig. 4. From Fig. 4a and Fig. 4b, under the assumption of 110 dB SIC capability for both network configurations, positive gains (56 percent and 16 percent for the uniform and clustered cases respectively) of FD networking in system SE can be achieved even when no extra interference management strategy is used. This is because the inter-cell interference leads to a smaller SE in each cell than that in the single-cell case in Fig. 2b. However, all bandwidth could be used for both UL and DL UEs, so user diversity helps maintain system throughput in the FD case. It is encouraging to see an extra 20 percent or 35 percent gain by applying single-cell based power control or UE pairing on top for the uniform case and clustered case, respectively. This verifies the earlier observation that the intra-cell inter-UE interference is most dominant under our setting. From the figure, the gain for the clustered case is higher because the intra-cell inter-UE interference is more severe, as in Fig. 3. Moreover, an extra 56 percent gain for the clustered case can be obtained when these two interference management strategies are combined, which implies that the FD networks will perform interference-aware UE pairing and even fall back to HD mode to ensure no performance degradation.

Next, we discuss how the SE gain of the FD networks over the HD networks could be with different SIC capabilities. Figure 4c shows for the standard PF scheduling method in the uniform case, it needs at least an 83 dB SIC capability to achieve the sum rate gain of the FD networks, and requires a less effective SIC capability if better interference management schemes are leveraged. Moreover, with power control,

It is encouraging to see an extra 20 percent or 35 percent gain by applying single-cell based power control or UE pairing on top for the uniform case and clustered case, respectively. This verifies the earlier observation that the intra-cell inter-UE interference is most dominant under our setting.

For both configurations, the FD networks with the standard PF scheduling method could yield larger EE (24 percent and 4 percent for the uniform and clustered cases respectively) than the HD networks. By exploiting the power control and the UE pairing methods, the EE performance improvement could be as large as 110 percent.

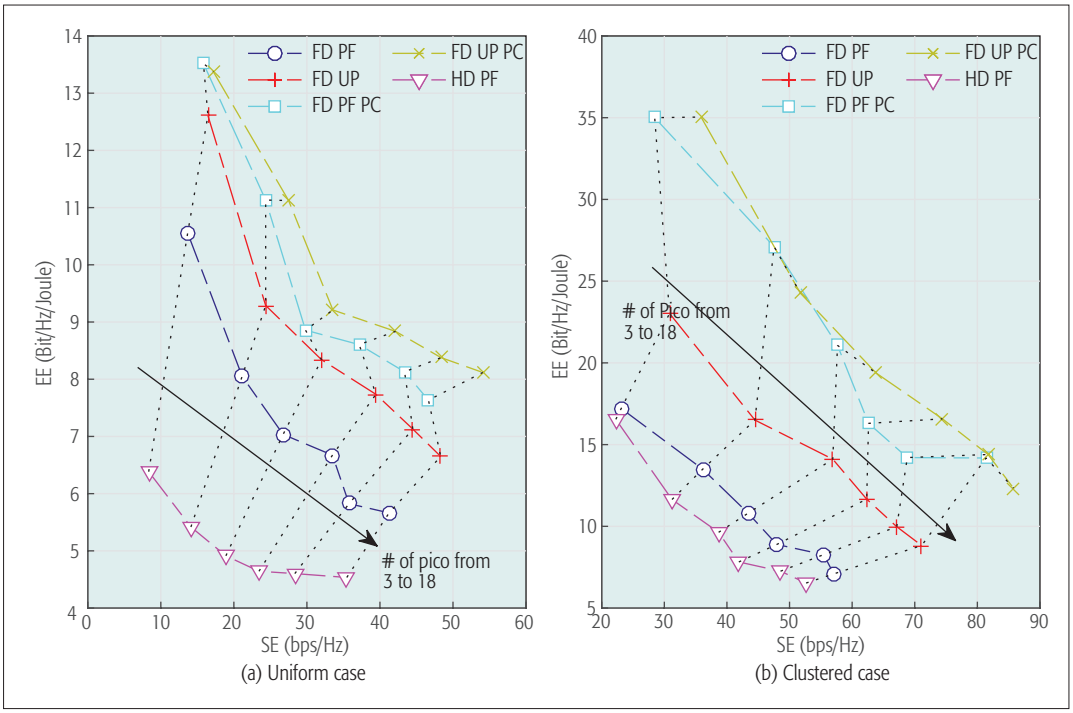


Figure 5. The SE and EE performance of FD cellular networks: a) uniform case with 192 UEs per macro BS; b) clustered case with 8 clustered UEs per pico BS.

Key Findings: (1) FD networks yield similar SE-EE trend but better tradeoff curves to HD networks. (2) By exploiting power control and UE pairing schemes, the EE performance improvement could be as large as 110%.

the FD networks could fall back to HD mode whenever necessary to reap a larger SE at some transmission direction (i.e., UL and DL). Hence, it always exhibits performance improvement even when the SIC capability is not so effective. On the other hand, Fig. 4d shows that stronger SIC capability, around 100 dB, is needed to mitigate the negative impact of other kinds of interference in the clustered case, such as the intra-cell inter-UE interference.

SE-EE Relationship

Figure 5 further presents the system SE and EE performance of the FD networks. From the figures, there are similar SE-EE tradeoff curves in both network configurations. When the number of pico BSs per macro BS increases from three to 18, it leads to distinct variation trends in SE and EE, because more pico BSs imply a higher frequency reuse ratio and thus lead to a larger SE. However, deploying more pico BSs also adds to the total power consumption and incurs larger inter-cell interference. Consequently, the SE gain cannot compensate for the loss in interference and power consumption, resulting in the EE decrease. Meanwhile, in addition to the benefit to SE performance improvement already validated in Fig. 4, the FD networks could also benefit system EE, as shown in Fig. 5. For both configurations, the FD networks with the standard PF scheduling method could yield larger EE (24 percent and 4 percent for the uniform and clustered cases respectively) than the HD networks. By exploiting the power control and the UE pairing methods, the EE performance improvement could be as large as 110 percent.

Conclusions and Future Works

From the discussion in this article, we found that equipping pico BSs with FD functionality will be most practical and promising for FD communications in cellular networks. Starting with a single-cell FD network, we discovered that the power control solution for any given UL and DL UE pair has a binary feature, and thus the system SE optimization problem reduces to a UE pairing problem. We further demonstrated the importance of interference-awareness in pairing UEs. For the multi-cell scenario, our interference analysis results showed intra-cell inter-UE interference is most dominant under our setting. Therefore, we further combined the UE pairing scheme based on distance-aware joint PF scheduling and the binary power control scheme as the interference management solution for the multi-cell FD networks. The system-level simulation has proven up to 91 percent and 72 percent SE gains over the traditional HD networks for the uniform and clustered cases, respectively, under 110 dB SIC capability. Therefore, we could conclude that FD works for cellular networks!

However, there still exist demanding challenges to address, including the combination of FD functionality with multiple-input mutiple-output (MIMO) systems, the protocol and algorithm design to take advantage of interference cancellation at the receiver or even to combine with the non-orthogonal multiple access schemes [15], as well as the extension to UEs with FD capability in both cellular and device-to-device (D2D) communications.

Acknowledgment

We want to express our sincere thanks to Eddy Hum, Huan Wu, and Moshiur Rahman from Huawei Canada for their insightful comments to improve the quality of this article. This work is supported by the National Science and Technology Major Project of China (No. 2015ZX03002010). R. Li's work is also in part supported by the National Postdoctoral Program for Innovative Talents of China (No. BX201600133).

References

[1] S. Hong *et al.*, "Applications of Self-Interference Cancellation in 5G and Beyond," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 114–21.
[2] D. Bharadia, E. McMilin, and S. Katti, "Full Duplex Radios," *Proc. ACM SIGCOMM 2013*, Hong Kong, China, Aug. 2013.
[3] M. Duarte and A. Sabharwal, "Full-Duplex Wireless Communications Using Off-the-Shelf Radios: Feasibility and First Results," *Proc. ASILOMAR 2010*, Pacific Grove, CA, Nov. 2010.
[4] Huawei, "Full-Duplex Technology for 5G," *Huawei Innovation Research Program J.*, pp. 62–69, Jan. 2015, http://www-file.huawei.com/~/media/CORPORATE/PDF/Downloads/Inaugural-Issue-5G-Research-and-Innovation.pdf
[5] M. Chung *et al.*, "Prototyping Real-Time Full Duplex Radios," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 56–63.
[6] Z. Shen *et al.*, "Dynamic Uplink-Downlink Configuration and Interference Managemen in TD-LTE," *IEEE Commun. Mag.*, vol. 50, no. 11, Nov. 2012, pp. 51–59.
[7] S. Goyal *et al.*, "Full Duplex Cellular Systems: Will Doubling Interference Prevent Doubling Capacity?" *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 121–27.
[8] R. A. Sultan *et al.*, "Mode Selection, User Pairing, Subcarrier Allocation and Power Control in Full-Duplex OFDMA hetnets," *Proc. IEEE ICC 2015 (Small-Nets WKSP)*, London, UK, June 2015.
[9] A. Sabharwal *et al.*, "In-Band Full-Duplex Wireless: Challenges and Opportunities," *IEEE JSAC*, vol. 32, no. 9, Sept. 2014, pp. 1637–52.
[10] Y. Xin *et al.*, "Co-Channel Interference Suppression Techniques for Full Duplex Cellular System," *China Commun.*, vol. 12, no. Supplement, Dec. 2015, pp. 18–27.
[11] S. Han *et al.*, "Full Duplex: Coming into Reality in 2020?" *Proc. IEEE Globecom 2014*, Austin, TX, USA, Dec. 2014.
[12] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-Driven Characterization of Full-Duplex Wireless Systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4296–4307, Dec. 2012.
[13] S. Shao *et al.*, "Analysis of Carrier Utilization in Full-Duplex Cellular Networks by Dividing the Co-Channel Interference Region," *IEEE Commun. Lett.*, vol. 18, no. 6, June 2014, pp. 1043–46.
[14] 3GPP, "Further Enhancements to LTE Time Division Duplex (TDD) for Downlink-Uplink Interference Management and Traffic Adaptation (TS 36.828)," Sept. 2012, http://www.3gpp.org/dynareport/36828.htm
[15] L. Lu *et al.*, "Prototype for 5G New Air Interface Technology SCMA and Performance Evaluation," *China Commun.*, vol. 12, no. Supplement, Dec. 2015, pp. 38–48.

Biographies

Rongpeng Li [S'12, M'17] (lirongpeng@zju.edu.cn) received his Ph.D and B.E. degrees from Zhejiang University, Hangzhou, China and Xidian University, Xi'an, China in June 2015 and June 2010, respectively, both as "Excellent Graduates". He is now a postdoctoral researcher at the College of Computer Science and Technologies and College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. From August 2015 to September 2016, he was a researcher at the Wireless Communication Laboratory, Huawei Technologies Co. Ltd., Shanghai, China. He was a visiting doctoral student in Supélec, Rennes, France from September 2013 to December 2013, and an intern researcher at the China Mobile Research Institute, Beijing, China from May 2014 to August 2014. His research interests currently focus on applications of artificial intelligence, data-driven network design, and resource allocation of cellular networks (especially full-duplex networks). He was granted by the National Postdoctoral Program for Innovative Talents, which had a grant ratio of 13 percent in 2016. He is an editor of *China Communications*.

Yan Chen (bigbird.chenyan@huawei.com) received her B.Sc. and Ph.D. degrees in 2004 and 2009, respectively, from Chu Kochen Honored College, Zhejiang University, and the Institute of Information and Communication Engineering, Zhejiang University, respectively. She was a visiting researcher at the University of Science and Technology (HKUST) from 2008 to 2009. In the same year of graduation, she joined Huawei Technologies (Shanghai) Co., Ltd.. She was the team leader and project manager of the internal project Green Radio Excellence in Architecture and Technology (GREAT) from 2010 to 2013, during which she was also the project leader of the umbrella project Green Transmission Technologies (GTT) with the GreenTouch™ Consortium. From 2013 she has been the technical leader and project manager of the internal 5G air interface design project focusing on new waveform, grant-free and non-orthogonal multiple access, flexible duplex, advanced receiver, as well as communication system design toward ultra low latency and ultra high reliability performance. Her current research interests are more toward future communication system design to efficiently support multiplexing of different service scenarios with diversified requirements, in which new technologies from cross-layer optimization, control theory, and artificial intelligence need to be jointly exploited.

Geoffrey Ye Li [S'93, M'95, SM'97, F'06] (liye@ece.gatech.edu) is a full professor with the School of Electrical and Computer Engineering at the Georgia Institute of Technology as an associate professor and then a full professor. He also holds a Cheung Kong Scholar title at the University of Electronic Science and Technology of China since 2006. His general research interests include statistical signal processing and communications, with an emphasis on cross-layer optimization for spectral- and energy-efficient networks, cognitive radios and opportunistic spectrum access, and practical issues in LTE systems. In these areas, he has published around 400 papers in various journals and conferences in addition to 26 granted patents. His publications have been well cited from Google Citations, and he has been recognized as a Highly-Cited Researcher by Thomson Reuters. He has received several paper and/or achievement awards. Recently, he received the 2015 Distinguished Faculty Achievement Award from the School of Electrical and Computer Engineering, Georgia Tech.

Guangyi Liu (liuguangyi@chinamobile.com) received his B.S. in physics from Chinese Ocean University in 1997, and received his M.S. and Ph.D. degrees in circuits and systems from Beijing University of Posts and Telecommunications (BUPT) in 2000 and 2006, respectively. Since 2006, he has been working for the Research Institute of China Mobile Communication Corporation. He is currently the Chief Technical Officer of the wireless department at the China Mobile Research Institute (R&D of China Mobile), where he is in charge of wireless technology R&D, including LTE/LTE-Advanced and 5G. He has filed more than 80 patents. He is a very distinguished leader of the mobile communication industry and active in global industrialization, e.g., acting as the chair of the spectrum working group of GTI (Global TD-LTE Initiative), and vice chair of the CCSA TC5 WG6. Before he joined China Mobile in 2006, he worked for three years at Siemens and Shanghai Bell (now ALU) on 3G R&D.

The system-level simulation has proven up to 91 percent and 72 percent SE gains over the traditional HD networks for the uniform and clustered cases, respectively, under 110 dB SIC capability. Therefore, we could conclude that FD works for cellular networks!

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

EDUCATION & TRAINING: SCHOLARSHIP OF TEACHING AND SUPERVISION

BACKGROUND

The Scholarship of Teaching and Learning (SoTL) encourages educators to examine their own classroom practice, record their successes and failures, and ultimately share their experiences in a formal and scholarly way so that others may reflect on their findings and build upon teaching and learning processes.

SoTL acknowledges that concerns for privacy and other ethical issues associated with studies involving human subjects place limits on the types of research that can be conducted in classroom setting. Nevertheless, SoTL provides a mechanism for raising the standard of discussion concerning teaching and learning in the literature.

The Scholarship of Research and Supervision (SoRL) is a related concept that invites the same reflective approach to improving the quality of training through research, especially that conducted at the postgraduate level. SoRL invites researchers to examine their own supervisory practice, record their successes and failures, and ultimately share their experiences in a formal and scholarly way so that others may reflect on their findings and improve research and supervision processes.

This feature topic on the Scholarship of Teaching and Supervision is intended to hasten the incorporation of SoTL and SoRL into communications engineering curricula by providing educators and researchers with an opportunity to share their experience, best practices and case studies.

SCOPE OF SUBMISSIONS

Authors from industry, government and academia are invited to submit papers for this FT (Feature Topic) of IEEE Communications Magazine on the Scholarship of Teaching and Supervision. The FT scope includes, but is not limited to the following topics of interest:

- SoTL-based case studies of communications engineering curricula.
- Best practices for conducting SoTL-based studies in communications engineering curricula.
- SoTL-based case studies of professional training in communications engineering.
- Best practices for conducting SoTL-based studies in professional training in communications engineering.
- SoRL-based case studies of research-based training in communications engineering.
- Best practices for conducting SoRL-based studies of research-based training in communications engineering.
- Development of tools for conducting SoTL and SoRL-based studies.

SUBMISSIONS

Articles should be tutorial in nature, with the intended audience being all members of the communications technology communities. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words. Figures and tables should be limited to a combined total of six. The number of archival references is not to exceed 15. Complete guidelines for manuscript preparation can be found via the link http://www.comsoc.org/commag/paper-submission-guidelines. Please send a PDF (preferred) or MS-Word formatted paper via Manuscript Central (http://mc.manuscript-central.com/commag-ieee). Register or log in, and go to the Author Center. Follow the instructions there. Select "November 2017 / Scholarship of Teaching and Supervision."

IMPORTANT DATES

Manuscript Submission Deadline: May 1, 2017

Decision Notification: July 15, 2017

Final Manuscript Due Date: August 15, 2017

FT Publication Date: November 2017

GUEST EDITORS

David G Michelson (Lead)
Univ. of British Columbia, Canada
davem@ece.ubc.ca

Peter Ostafichuk
Univ. of British Columbia, Canada
Email: ostafich@mech.ubc.ca

C. Kelly Ottman
Milwaukee School of Engineering, USA
Email: ottman@msoe.edu