July 2015, Vol. 53, No. 7

IEEE Communications MAGAZINE

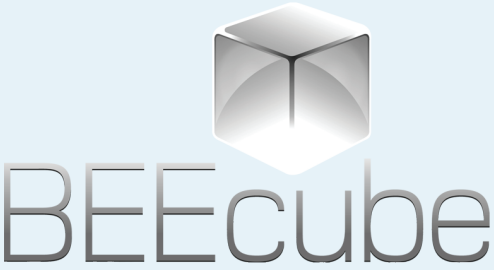
www.comsoc.org

- 5G Spectrum: Enabling the Future Mobile Landscape
- Network and Service Management
- Ad Hoc and Sensor Networks
- Cellular Cognitive Systems

A Publication of the IEEE Communications Society

IEEE Communications MAGAZINE

THANKS OUR CORPORATE SUPPORTERS



July 2015, Vol. 53, No. 7

IEEE Communications MAGAZINE

www.comsoc.org

- 5G Spectrum: Enabling the Future Mobile Landscape
- Network and Service Management
- Ad Hoc and Sensor Networks
- Cellular Cognitive Systems

A Publication of the IEEE Communications Society

Your 5G Eureka moment will happen sooner or later.

We'll help make it sooner.

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

HARDWARE + SOFTWARE + PEOPLE = 5G INSIGHTS

PEOPLE

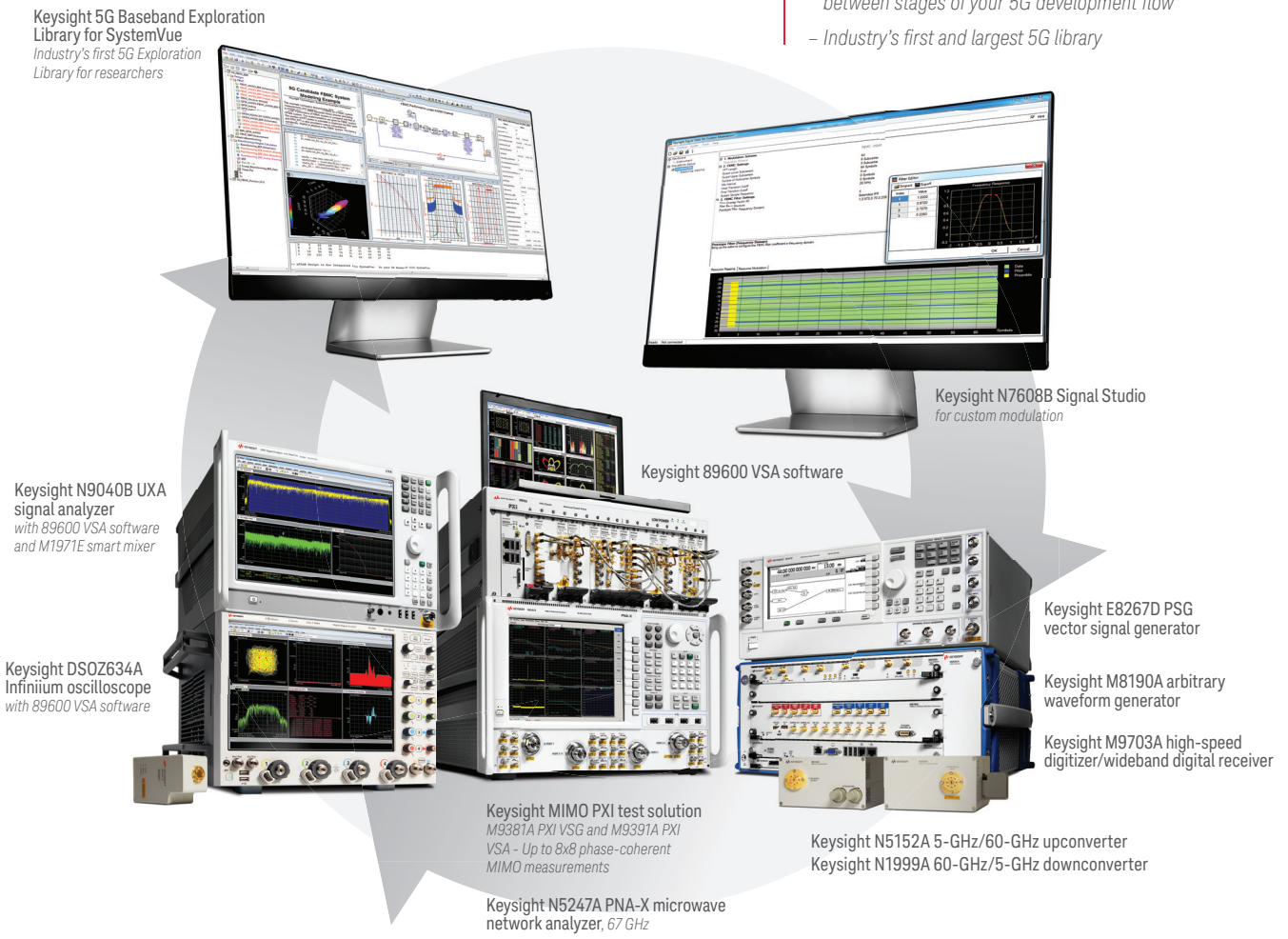
- *Keysight engineers are active in the leading 5G forums and consortia*
- *Keysight engineers are keynote speakers at 5G conferences and key contributors in top technical journals*
- *Applications engineers are in more than 100 countries around the world*

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

USA: 800 829 4444 CAN: 877 894 4414

HARDWARE + SOFTWARE

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



KEYSIGHT TECHNOLOGIES

Unlocking Measurement Insights

IEEE Communications Magazine

JULY 2015, Vol. 53, No. 7

www.comsoc.org/commag

6 **The President's Page**
9 **Conference Calendar**
11 **Global Communications Newsletter**
15 **ICIN 2015: Innovations in Services, Networks and Clouds**
208 **Advertisers' Index**

5G Spectrum: Enabling the Future Mobile Landscape

Guest Editors: Hans D. Schotten, Mikko A. Uusitalo, Jose F. Monserrat, and Olav Queseth

16 **Guest Editorial**
18 **Spectrum Access System for the Citizen Broadband Radio Service**
Munawwar M. Sohul, Miao Yao, Taeyoung Yang, and Jeffrey H. Reed
26 **Toward Spectrum Sharing: Opportunities and Technical Enablers**
Konstantinos Chatzikokolakis, Panagiotis Spapis, Alexandros Kaloxylos, and Nancy Alonistioti
34 **Coordination Protocol for Inter-Operator Spectrum Sharing in Co-Primary 5G Small Cell Networks**
Bikramjit Singh, Sofonias Hailu, Konstantinos Koufos, Alexis A. Dowhuszko, Olav Tirkkonen, Riku Jäntti, and Randall Berry
42 **Spectrum and License Flexibility for 5G Networks**
Adrian Kliks, Oliver Holland, Arturo Basaure, and Marja Matinmikko
50 **Broadcast Television Spectrum Incentive Auctions in the U.S.: Trends, Challenges, and Opportunities**
David Gómez-Barquero and M. Winston Caldwell
58 **5G Spectrum: Is China Ready?**
Tan Wang, Gen Li, Jiaxin Ding, Qingyu Miao, Jingchun Li, and Ying Wang

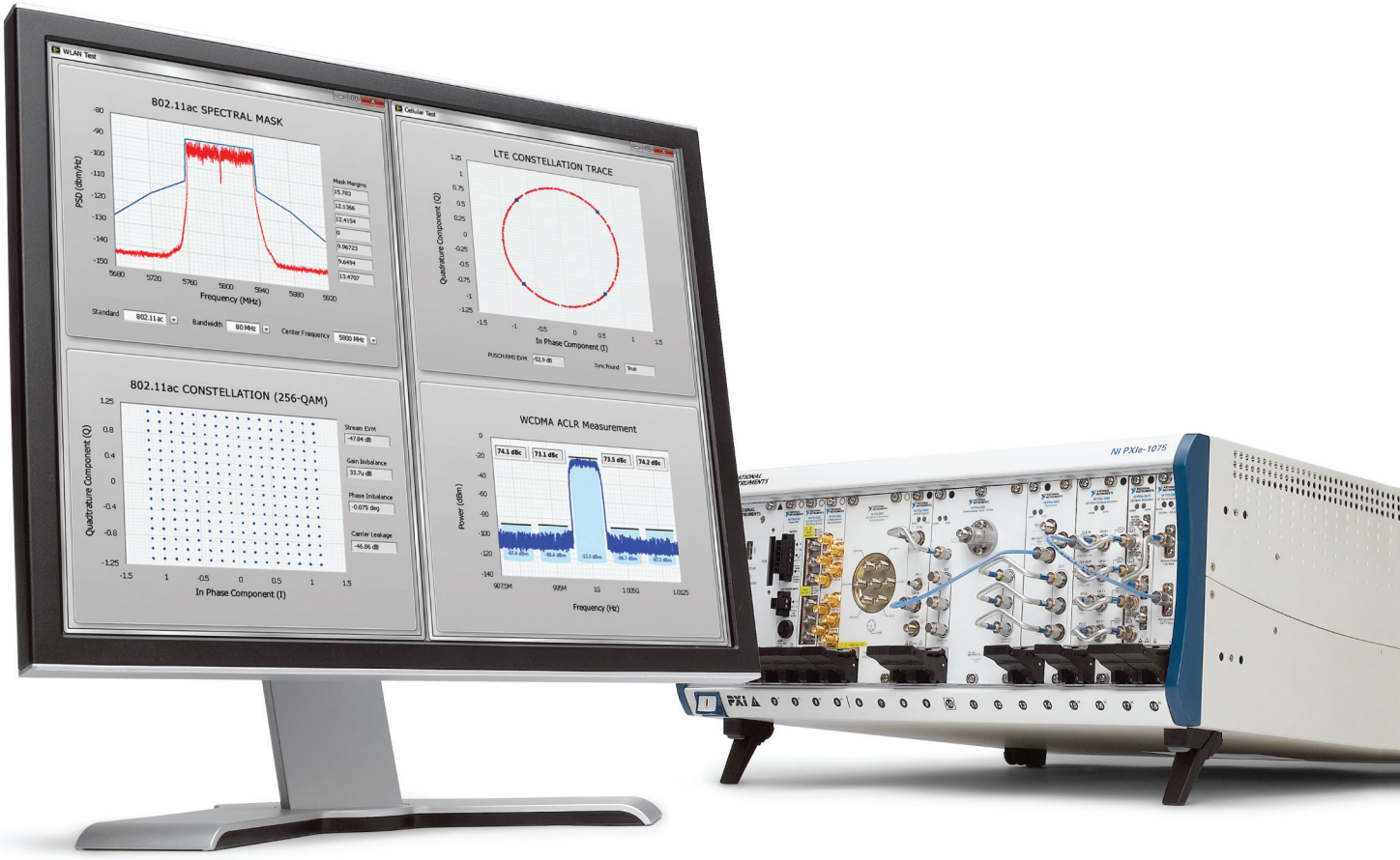
Emerging Applications, Services, and Engineering for Cellular Cognitive Systems: Part II

Guest Editors: Muhammad Zeeshan Shakir, Octavia A. Dobre, Muhammad Ali Imran, Apostolos Papathanassiou, Zhongshan Zhang, Athanasios V. Vasilakos, Honggang Wang, and Hiroshi Harada

66 **Guest Editorial**
70 **Solar Energy Empowered 5G Cognitive Metro-Cellular Networks**
Syed Ali Raza Zaidi, Asma Afzal, Maryam Hafeez, Mounir Ghogho, Desmond C. McLernon, and Ananthram Swami
78 **Virtualized Cognitive Network Architecture for 5G Cellular Networks**
Hesham ElSawy, Hayssam Dahrouj, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouini
86 **Enhanced Multi-Parameter Cognitive Architecture for Future Wireless Communications**
Feifei Gao and Kaiqing Zhang
93 **CPC-Based Backward-Compatible Network Access for LTE Cognitive Radio Cellular Networks**
Ling Liu, Yiqing Zhou, Lin Tian, and Jinglin Shi
100 **A Game-Theoretic Perspective on Self-Organizing Optimization for Cognitive Small Cells**
Yuhua Xu, Jinlong Wang, Qihui Wu, Zhiyong Du, Liang Shen, and Alagan Anpalagan

Redefining RF and Microwave Instrumentation

with open software and modular hardware



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

WIRELESS TECHNOLOGIES

National Instruments supports a broad range of wireless standards including:

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

>> Learn more at ni.com/redefine

800 813 5078

NATIONAL INSTRUMENTS™

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

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

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

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

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

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

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

109 **COGNITIVE VEHICULAR COMMUNICATION FOR 5G**
SHAHID MUMTAZ, KAZI MOHAMMED SAIDUL HUQ, MUHAMMAD IKRAM ASHRAF, JONATHAN RODRIGUEZ, VALDEMAR MONTEIRO, AND CHRISTOS POLITIS

118 **COGCELL: COGNITIVE INTERPLAY BETWEEN 60 GHZ PICOCELLS AND 2.4/5 GHZ HOTSPOTS IN THE 5G ERA**
KISHOR CHANDRA, R. VENKATESHA PRASAD, BIEN QUANG, AND I. G. M. M. NIEMEGEERS

126 **COGNITIVE SPECTRUM ACCESS IN DEVICE-TO-DEVICE-ENABLED CELLULAR NETWORKS**
AHMED HAMDI SAKR, HINA TABASSUM, EKRAM HOSSAIN, AND DONG IN KIM

NETWORK AND SERVICE MANAGEMENT
SERIES EDITORS: GEORGE PAVLOU AND JÜRGEN SCHÖNWÄLDER

134 **SERIES EDITORIAL**

136 **TOWARD A HOLISTIC FEDERATED FUTURE INTERNET EXPERIMENTATION ENVIRONMENT: THE EXPERIENCE OF NOVI RESEARCH AND EXPERIMENTATION**
V. MAGLARIS, C. PAPAGIANNI, G. ANDROULIDAKIS, M. GRAMMATIKOU, P. GROSSO, J. VAN DER HAM, C. DE LAAT, B. PIETRZAK, B. BELTER, J. STEGER, S. LAKI, M. CAMPANELLA, AND S. SALLENT

145 **QUALITY OF EXPERIENCE MANAGEMENT IN MOBILE CELLULAR NETWORKS: KEY ISSUES AND DESIGN CHALLENGES**
EIRINI LIOTOU, DIMITRIS TSOLKAS, NIKOS PASSAS, AND LAZAROS MERAKOS

154 **AN OPEN FRAMEWORK FOR PROGRAMMABLE, SELF-MANAGED RADIO ACCESS NETWORKS**
KOSTAS TSAGKARIS, GEORGE POULIOS, PANAGIOTIS DEMESTICHAS, ABDOULAYE TALL, ZWI ALTMAN, AND CHRISTIAN DESTRÉ

162 **ON-DEMAND SCHEDULING: ACHIEVING QOS DIFFERENTIATION FOR D2D COMMUNICATIONS**
MIN SHENG, HONGGUANG SUN, XIJUN WANG, YAN ZHANG, TONY Q. S. QUEK, JUNYU LIU, AND JIANDONG LI

AD HOC AND SENSOR NETWORKS
SERIES EDITORS: EDOARDO BIAGIONI AND SILVIA GIORDANO

171 **SERIES EDITORIAL**

172 **PEDESTRIAN MOBILITY IN THEME PARK DISASTERS**
GÜRKAN SOLMAZ AND DAMLA TURGUT

ACCEPTED FROM OPEN CALL

178 **ON THE LIMITS OF PREDICTABILITY IN REAL-WORLD RADIO SPECTRUM STATE DYNAMICS: FROM ENTROPY THEORY TO 5G SPECTRUM SHARING**
GUORU DING, JINLONG WANG, QIHUI WU, YU-DONG YAO, RONGPENG LI, HONGGANG ZHANG, AND YULONG ZOU

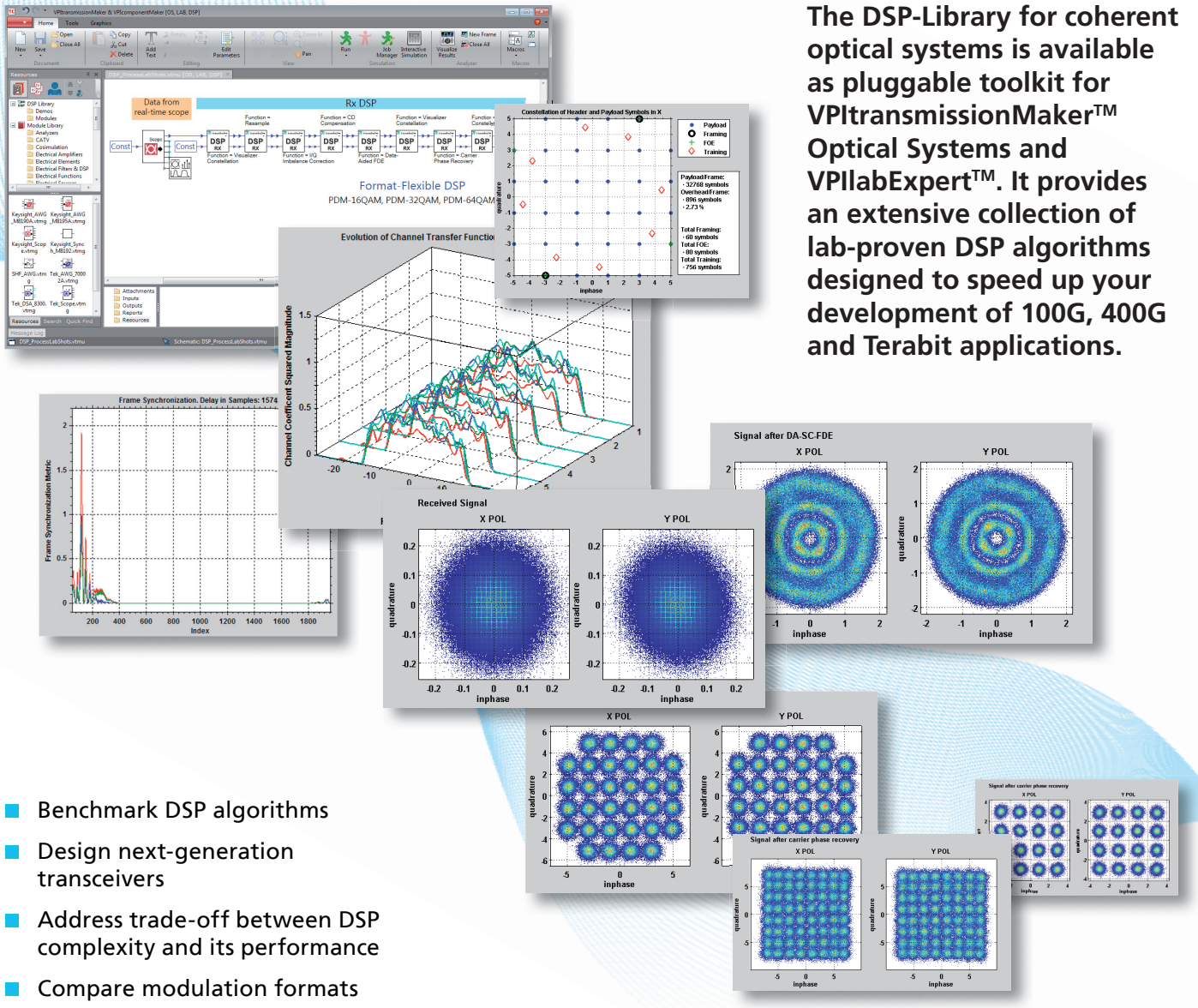
184 **AN EVOLUTIONARY PATH FOR THE EVOLVED PACKET SYSTEM**
MARC PORTOLES-COMERAS, JOSEP MANGUES-BAFALLUY, ANDREY KRENDZEL, MANUEL REQUENA-ESTESO, AND ALBERT CABELLOS-APARICIO

192 **COVERAGE ENHANCEMENT TECHNIQUES FOR MACHINE-TO-MACHINE COMMUNICATIONS OVER LTE**
GHASEM NADDAFZADEH-SHIRAZI, LUTZ LAMPE, GUSTAV VOS, AND STEVE BENNETT

201 **LISP: A SOUTHBOUND SDN PROTOCOL?**
ALBERTO RODRIGUEZ-NATAL, MARC PORTOLES-COMERAS, VINA ERMAGAN, DARREL LEWIS, DINO FARINACCI, FABIO MAINO, ALBERT CABELLOS-APARICIO

Fraunhofer
Heinrich Hertz Institute

Ready-to-use DSP-Library for optical system simulations and experiments



The DSP-Library for coherent optical systems is available as pluggable toolkit for VPItransmissionMaker™ Optical Systems and VPIlabExpert™. It provides an extensive collection of lab-proven DSP algorithms designed to speed up your development of 100G, 400G and Terabit applications.

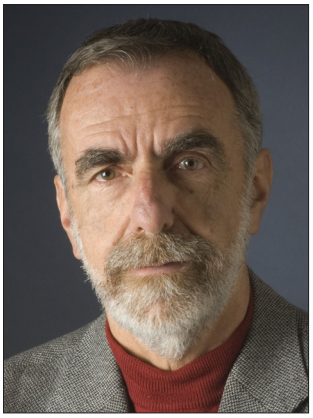
- Benchmark DSP algorithms
- Design next-generation transceivers
- Address trade-off between DSP complexity and its performance
- Compare modulation formats
- System performance analysis
- Define component requirements

In Cooperation with:

VPIphotonics
DESIGN AUTOMATION

Further Information:
http://www.vpiphotonics.com/Tools/OpticalSystems/Toolkits/

CQ COMSOC: CALLING ALL COMSOC RADIO AMATEURS



SERGIO BENEDETTO

DOUG ZUCKERMAN

The Communications Society is a community of professionals with an interest in the technologies and applications associated with communications. Amateur radio is a community of hobbyists with an interest in the technologies and applications associated with communications, in particular by radio. Indeed, for many, ham radio was the start of a personally satisfying lifelong career in communications. Doug Zuckerman, a past ComSoc president and ham with call sign W2XD, joins me in this month's column to remind us of our ham radio heritage and suggest ways to rebuild ties that may have been lost over the years.

An active volunteer for more than 30 years, Doug Zuckerman is a past IEEE Division III (Communications Technology) Director, was 2008-2009 President of the IEEE Communications Society, and previously held leadership positions in conferences, publications, and membership development. He received his B.S., M.S., and Eng.Sc.D degrees from Columbia University, USA, and is an IEEE Life Fellow. His professional experience, mainly at Bell Labs and Telcordia Technologies, USA, spans the operations, management, and engineering of emerging communications technologies, networks, and applications. His work heavily influenced early standards for management of telecommunications networks. Presently semi-retired, he is still active in standards as a representative to the Optical Internetworking Forum. Much of his professional life has been dedicated to IEEE activities. His service resulted in the following honors: IEEE Third Millennium Medal, the IEEE Communications Society's McLellan Award for meritorious service, its Conference Achievement Award, and the Salah Aidarous Memorial Award. As might be guessed, ham radio played an important role in setting his career course.

WHAT IS HAM RADIO

From the amateur radio organization, the ARRL (www.arrl.org): "Amateur Radio (Ham Radio) is a popular hobby and service in which licensed Amateur Radio operators (hams) operate communications equipment." Each ham has a unique call sign (e.g., W2XD), with prefixes and numbers assigned by ITU-R and national administrations based on geography and other considerations. Anyone can become a ham, but why would they do so? There are many reasons. For many, especially us old timers who began in the 1960s, it was being able to solder together components such as resistors and capacitors, fire up vacuum tube transmitters and receivers connected to a long piece of wire, and have two-way communication with another ham down the road or on the other side of the planet. In the 1960s, this communication was either by voice (Amplitude Modulation or Single Sideband) or continuous wave (CW, using Morse Code). Today's hams have a much wider variety of modulation formats to choose from. Indeed, some lament the decline in the use of CW as these newer and more efficient digital technologies evolve.

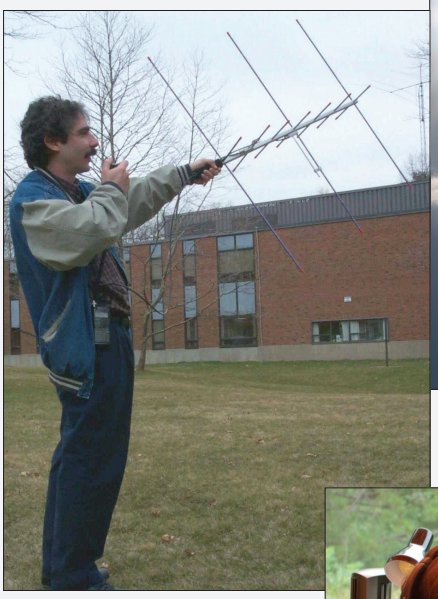
Ham radio actually has a number of subhobbies, summarized as follows:

Experimentation: Some hams like to build their own radio gear, be it a solid state packet switch, radio frequency transceiver, or an antenna design based on readily available antenna modeling software. These hams have a practical knowledge of communications theory and build their stations on that foundation augmented by trial and error.

Chasing DX (Distant Communications): Other hams like to communicate with stations in far off places, the further and more isolated the better. To help fellow DX-ers contact rare locations, some hams even go on "DXpeditions" to set up stations in sparsely populated regions of the world. South Pacific reef islands are a favorite. Confirmation cards, known as "QSL cards," are sent afterward to provide proof that the contacts took place. Various awards are available to recognize those who make a large number of contacts, e.g., the "DX Century Club Award" for those who've contacted at least 100 different countries. A variation is that some hams try to use as little power as possible to communicate (this is called, "QRP"), while others use the full legal power limit. The best DXers have a keen understanding of radio wave propagation, and pick frequencies accordingly.

Contesting: Even in a hobby, it can be all about winning. Throughout the year, some hams compete with each other in a variety of radio communications contests. Typically, these contests involve making as many contacts as possible over a pre-defined time period. The ham with the greatest number of contacts wins the contest. In some parts of the world, contesting is known as "Radiosport." Contests and radiosport may involve either individuals or teams. The goal of being best in these contests even drives experimentation to build highly competitive ham radio stations, and chasing DX improves operating skills and efficiency, both relevant to doing well in a contest.

Emergency Communications: This is where hams show their true value to society by serving the public during times of emergency. Because ham radio communications capabilities are highly distributed and independent of each other, communications is possible even when more centralized commercial facilities are disrupted by disasters. Hams active in emergency communications follow strict procedures and hone their skills by providing communications in non-disaster events such as marathons.



Handheld experimentation.

ND5P antenna suitable for DX and contests.

K1SEZ — providing marathon communications.

N2NT at his station.

Traffic Handling: Traffic handling, organized in the U.S. through the National Traffic System (other countries have similar systems) in the 1950s, provided a free "radiogram" service to the public. Through a hierarchical structure of local, section, region, and area "nets" interleaved to meet at the same time each day, hams would use a well-defined form and precise procedures to manually relay messages from one part of the country to another (or even to other countries). The messages had to be non-commercial, such as birthday greetings. The system was very popular at a time when a phone call to another part of town wasn't free, and it was way too costly to call outside the local area. It was the era of people communicating through mailing hand written letters to each other. Traffic handling was like a hands-on very slow packet network, where the radiogram is the packet, and each traffic "net" is the packet switch. Though we now can call or email anywhere in the world for pennies (or less), one would think there's no longer a need for traffic handling. Yet, the National Traffic System still exists, and many hams still actively handle traffic using CW, voice, or even packet radio.

PERSONAL EXPERIENCES

Speaking with hams who are also active in IEEE or ComSoc shows that they have similar stories about their early ham radio pursuits and long-term career impact. As an example, I was first licensed in 1961 when I was 14 years old. My first station was a Heathkit 50-watt crystal controlled CW transmitter, super-heterodyne all band receiver with vacuum tubes, and a 40-meter band dipole antenna on the roof of our four-story walk-up apartment building in Brooklyn, New York. Television interference was a big problem in those days, and I had to regularly replace my wire antennas after the neighbors cut them down. My first contact with another ham in Brooklyn was amazing, but not as amazing as when I made my first DX contact with a ham in Venezuela. I was hooked!

I continued being active through high school, but went "dormant" during my college years starting in 1965 at Columbia (which had a ham radio club, by the way). Studies took up almost all my time. Ham radio made electrical engineering a simple choice. Since I had a very strong interest in antennas (remember all the antennas I had experimented with and had to replace upon removal by the neighbors), I focused my masters and doctorate work on electromagnetics. This set the basis for being hired in 1969 by Bell Labs in Holmdel, New Jersey, by a group working on a long-haul millimeter-wave transmission system (40-100 GHz), then satellite antennas and systems, and then operations and maintenance of digital transmission systems, AT&T digital network planning, standards activities, etc.

In my group at Bell Labs, most of my colleagues were

hams, some of whom were very active on the airwaves. This reignited my interest and I got back into the hobby. "Ham radio is a way of life!" became our slogan. Since I never had a decent antenna nor high power gear, my main focus in ham radio was on traffic handling. The ham radio community actually has a governance structure, and I volunteered my time to help lead and manage a large portion of the National Traffic System, as well as serve as an officer in Bell Labs' Holmdel Amateur Radio Club. I continued being way too active until it became, "Enough, already!" At that point, the hobby went dormant again, though my call sign lives on through my email address, w2xd@aol.com, and I still enjoy reading *QST Magazine*, the monthly ham radio magazine that I receive as part of a life membership in the ARRL

Much of what I experienced and learned through ham radio carried over to my earlier Bell Labs career and my participation in IEEE and ComSoc activities. I believe that many IEEE members can say ham radio played a key role in setting them on their way to a long and satisfying professional career.

CONNECTING THE COMSOC HAMS

Without formal data, it's hard to estimate how many ComSoc members also are or were hams. Years ago, our flagship ICC and GLOBECOM conferences used to invite the local amateur radio clubs to organize special sessions for ham conference attendees. These sessions were free of charge, which is typical for ham radio events. Over time, these sessions have ceased in order to give more room for regular technical paper sessions. Nonetheless, there are ways we may wish to consider for revitalizing our relationship with this important community. Some ideas are as follows:

- Establish an on-line community at comsoc.community.org where ComSoc (and other) hams who have common interests may communicate (this would be a modern version of a ham radio "net" from the old days).
- Create a special interest group that would organize conference sessions, facilitate articles for our periodicals, and serve as a foundation for ham radio communications contributions to ComSoc's wide range of technical activities.
- Hold meetings at conferences similar to the receptions held for Young Professionals and Chapter Chairs, e.g., have one to two hour gatherings at our major conferences with the help of local ham radio groups.
- Form "ComSoc Amateur Radio Club" to serve as a focal point for ComSoc ham activities, including organizing contesting and public service event support.
- Interact with ham organizations to negotiate joint activities with organizations such as the ARRL, which publishes "practitioner" relevant periodicals such as QST and QEX, and which has "hamfests" (like a convention/conference) throughout the year.

Other IEEE organizations have already been acknowledging and nurturing the ham radio connection. For example, the Microwave Theory and Techniques Society headlined its January/February 2015 issue of *IEEE Microwave Magazine*, "Spreading the Word — Amateur Radio Operators Since the Beginning of EM Communications." Several years ago, the Consumer Electronics Society featured an editorial on ham radio in its newsletter. Harking back to this article's opening: "The Communications Society is a community of professionals with an interest in the technologies and applications associated with communications. Amateur radio is a community of hobbyists with an interest in the technologies and applications associated with communications, in particular by radio." Let's start thinking of ways we might use this overlap to enhance the value of being a ham and being in ComSoc.

OMBUDSMAN

COMSOC BYLAWS ARTICLE 3.8.10

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

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

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

2015

OCTOBER

LANOMS 2015 — Latin American Network Operations and Management Symposium, 1– 3 Oct.
Joao Pessoa, Brazil
http://www.lanoms.org/2015/

IEEE CLOUDNET 2015 — 4th IEEE Int'l. Conference on Cloud Networking, 5–7 Oct.
Niagara Falls, Canada
http://www.ieee-cloudnet.org/

RNDM 2015 — 7th Int'l. Workshop on Reliable Networks Design and Modeling, 5–7 Oct.
Munich, Germany
http://www.rndm.pl/2015/

WMNC 2015 — 8th IFIP Wireless and Mobile Networking Conference, 5–7 Oct.
Munich, Germany
http://www.wmnc2015.com/

ATC 2015 — Int'l. Conference on Advanced Technologies for Communications, 14–16 Oct.
Ho Chi Minh, Vietnam
http://www.rev-conf.org/

APCC 2015 — 21st Asia-Pacific Conference on Communications, 14–16 Oct.
Kyoto, Japan
http://www.apcc2015.ieice.org/

IEEE HEALTHCOM 2015, 17th IEEE Int'l. Conference on e-Health Networking, Application & Services, 14–17 Oct.
Boston, MA
http://www.ieee-healthcom.org/index.html

WCSP 2015 — Int'l. Conference on Wireless Communications & Signal Processing, 15–17 Oct.
Nanjing, China
http://www.ic-wcsp.org/

MILCOM 2015 — Military Communications Conference, 26–28 Oct.
Tampa, FL
http://events.jspargo.com/milcom15/public/enter.aspx

IOT 2015 — 5th Int'l. Conference on the Internet of Things, 26–28 Oct.
Seoul, Korea
http://www.iot-conference.org/iot2015/

CNSM 2015 — 11th Int'l. Conference on Standards for Communications and Networking, 26–30 Oct.
Barcelona, Spain
http://www.cnsm-conf.org/2015/

IEEE CSCN 2015 — IEEE Conference on Standards for Communications and Networking, 28–30 Oct.
Tokyo, Japan
http://www.ieee-cscn.org/

IEEE ICSOS 2015 — IEEE Int'l. Conference on Space Optical Systems and Applications, 27–28 Oct.
New Orleans, LA
http://icsos2015.nict.go.jp/

NOVEMBER

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

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

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

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

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

(Continued on next page)

–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.

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

DECEMBER

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

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

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

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

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

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

2016

JANUARY

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

MARCH

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

APRIL

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

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

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

ComSoc 2015 Election: Take Time to Vote

Ballots were e-mailed and/or postal mailed 29 May 2015 to all Higher Grade* IEEE Communications Society Members and Affiliates (excluding Students) whose memberships were effective prior to 1 May 2015. You must have an e-ballot or paper ballot before you can vote.

Vote Now using the URL below. You will need your IEEE Account username/password to access the ballot. If you do not remember your password, you may retrieve it on the voter login page.

https://eballot4.votenet.com/IEEE

If you have questions about the IEEE ComSoc voting process or would like to request a paper ballot, please contact ieee-comsocvote@ieee.org or +1 732 562 3904.

If you do not receive a ballot by 30 June, but you feel your membership was valid before 1 May 2015, you may e-mail ieee-comsocvote@ieee.org or call +1 732 562 3904 to check your member status. (Provide your member number, full name, and address.)

Please note IEEE Policy (Section 14.1) that IEEE mailing lists should not be used for "electioneering" in connection with any office within the IEEE.

Voting for this election closes 24 July 2015 at 4:00 p.m. EDT! Please vote!

*Includes Graduate Student Members

GCN GLOBAL COMMUNICATIONS NEWSLETTER

July 2015
ISSN 2374-1082

DISTINGUISHED LECTURE TOUR

DLT on Photonic Network of Tetsuya Kawanishi in Bandung, Indonesia

By Arief Hamdani Gunawan, Telkom Indonesia

On 21 February 2015, IEEE R10 Industry Relations initiated the IEEE Distinguished Lecture on Photonic Networks at Telkom University, Bandung. The activity that is also strongly supported by the IEEE Indonesia Section as well as the IEEE Communications Society Indonesia Chapter provides fertile conditions for engineering since government, industry, and academica support it.

The IEEE Distinguished Lecturer on Photonic Networks was Dr. Tetsuya Kawanishi, an IEEE Fellow and Director of the Lightwave Devices Laboratory, NICT. The other speakers are Dr. Denny Setiawan, Dr. Atsushi Kanno, Dr. Yusuf Nur Wijayanto, and Dr. Ismudiati Puri Handayani.

The first speaker, delivering the keynote, was Dr. Denny Setiawan, Head of the Fixed and Land Mobile Services Group, Directorate of Spectrum Policy and Planning, Directorate General of Resources and Standards, Ministry of Communications and IT, Republic of Indonesia. The topic was 'Accelerating Broadband Penetration in Indonesia.' The keynote also highlighted the importance of radio over fiber (RoF). After the keynote there were three topics related to activities of the National Institute of Information and Communications Technology (NICT), Japan: an overview of the Photonic Network Research Institute; seamless convergence of optical and radio technologies; and high functional electrical-to-optical conversion. There was also one topic from Telkom University.

Dr. Tetsuya Kawanishi, an IEEE Fellow and Director of the Lightwave Devices Laboratory, NICT, discussed the progress of photonics technology to realize high-speed and high-precision signal transport for future networks. Actually, future high-speed communication technology requires drastic improvement of the preciseness of optical signal generation and detection; advanced optical modulation techniques with higher order multi-level modulation is relayed on the precise optical modulator. High extinction ratio optical modulator with an extinction ratio up to 70 dB, which was developed by the NICT, is capable of advanced optical networking as well as precise clock distribution for radio astronomy and high-precision radar system for surveillance of airport runways.

Dr. Atsushi Kanno, Senior Researcher at the laboratory, showed a possible solution for realization of seamless convergence of optical and radio networks using radio over fiber (RoF) technology. The seamless conversion helps realize functionality of the link including possible reduction of transmission latency in the entire link. RoF technology also realizes direct signal conversion between the optical and radio signals in the millimeter-wave band and even the terahertz wave band. Thus, high-speed wireless signal transmission with capacity comparable to that in advanced optical communication is realized by a coherent RoF technology, which is comprised of a high-speed digital signal processing diverted from the recent optical digital coherent technology and the RoF technique.

Dr. Tetsuya Kawanishi, IEEE Fellows and Director of the Lightwave Devices Laboratory, NICT delivered the progress of the photonics technology to realize high-speed and high-precise signal transport for future network.

Dr. Yusuf Nur Wijayanto, a researcher at the laboratory, presented the high-speed electrical-to-optical signal converter realized by a lithium niobate optical modulator directly equipped with antennas on the substrate. In general, the optical modulator connected with the antennas has relatively large loss by connecting the transmission lines between the electrode of the modulator and the antenna, especially in high frequency radio signals. Dr. Wijayanto performed patch antenna array direct loaded on optical waveguides in the lithium niobate modulator. The technology is capable for direct conversion from high-speed and high-frequency radio signals to an optical signal. Additionally, the array antenna has the other function: to detect incident radio direction for applications to directional wireless communication, sensing, and radar.

Dr. Ismudiati Puri Handayani from Telkom University presented Ultrafast dynamics in TbMnO3. Demand on high speed data transmision in communication requires ultrafast optical switching, which involves compatible devices exhibiting ultrafast response time. In this respect, it is crucial to explore the dynamics of material response time under light illumination. The electron dynamics might involve other degrees of freedom, such as spin and lattice, which subsequently determine the characteristics of excited electron relaxation time. In the study, time resolved optical spectroscopy is used to elucidate the dynamics of photo excited electrons in TbMnO3, a multiferroic material with the Néel temperature TN,1 of 41 K and ferroelectric phase transtition, TN,2 of26 K. A 3,1 eV light, which corresponds to oxygen to manganite charge tranfer, is used to create free electrons while the 2 eV light is used to probe the transient responds. The fast response time is observed in the order of 30 ps. The result is interpreted as the relaxation of doped spin-aligned carriers in the presence of an underlying magnetic lattice in TbMnO3. The transient responses while probing d–d intersite transitions show marked differences along different crystallographic directions, which are discussed in terms of the interplay between the processes of hopping of the photo-injected electrons and the magnetic order in the material.

DISTINGUISHED LECTURE TOUR

DLT of Koichi Asatani to Montréal, Canada

By: Mouhamed Abdulla (U. Québec), Anader Benyamin-Seeyar (Concordia U.), and Fabrice Labeau (McGill U.)

On Feb. 18, 2015 we had the great pleasure to welcome Prof. Koichi Asatani to the IEEE Montréal Section in Canada, visiting us from his hometown in Kyoto, Japan. Prof. Koichi is no stranger to the IEEE ComSoc community, given his involvement in various leadership positions, the most recent one as the IEEE ComSoc Membership Programs Development (MPD) Director. In brief, he is an IEEE Fellow, IEICE Fellow, Professor Emeritus at Kogakuin U. (Tokyo, Japan), Chair Professor at Nankai U. (Tianjin, China), founder of the QoS Series Symposium at ICC and GLOBECOM, an Executive Chair for ICC'11 (Kyoto, Japan), and three times an IEEE ComSoc Distinguished Lecturer (DL): 2006-2009, 2011-2012, and 2013-2014. Moreover, he is a well-known speaker, and among other areas of proficiency in communications, he is an expert on Gigabit fiber optic networks, such as FTTH and G-PON.

Given the interest of our local membership in his research area and his expertise, it was only natural for us to invite Dr. Asatani to visit our Section. The idea then grew from a stand-alone DL event into a wider North American (NA) Tour (DLT), where Dr. Asatani would present his latest research in various cities across Eastern Canada and the U.S. With this idea in mind, we were pleased that the various NA ComSoc Chapters showed their interest in this initiative. But most importantly, we were delighted that Dr. Asatani was gracious enough to accept this multi-site traveling and lecturing commitment.

To begin his journey, Dr. Asatani left Tokyo on Feb. 17 and arrived on the same day in Montréal. The seminar was held on Feb. 18 from 6-8 pm EST at the Engineering Building of Concordia U. located in downtown Montréal. Despite the very cold temperature and heavy snow storm in Eastern Canadian, the event was a huge success, attracting more than 40 attendees. This is in part due to a strong advertising campaign on diverse platforms, including: IEEE vTools, IEEE eNotice Service, IEEE Montréal LinkedIn Group, Chapter membership drive, and of course Twitter. Moreover, personal emails were sent to local scholars, professionals, and graduate students via SYTACom, which is a collaborative center for telecommunications research in the province of Québec. The audience was diverse and composed of faculty members, post-doc researchers, and Ph.D. students from Concordia U., McGill U., U. Québec, U. Montréal, and INRS. We also had participants from the Canadian government (NRC-IRAP), and various national and multinational telecommunication companies, including Cisco Systems, Cogeco, and Fonex.

The title of Dr. Asatani's talk was "Trends and Issues of FTTH and G-PON" (slides: http://drmoe.org/ieee/KoichiTalk_2015.pdf). In short, Dr. Asatani succeeded in motivating the audience, by explaining the fundamentals, requirements, and regulatory aspects of Fiber-to-the-Home (FTTH) and Gigabit Passive Optical Networks (G-PON). He also made interesting remarks and predicted that FTTH broadband technology will eventually replace ADSL as the next-generation network due to its high-throughput range in the Gbps range. Furthermore, he discussed global developments, the future of FTTH and G-PON standardization activities, and technological trends. He answered many interesting questions, with exchanges and follow-up discussions, which were effectively answered by Prof. Asatani.

In addition to the technical component, from the start to the end of the talk, Dr. Asatani made the seminar friendlier by sharing some cultural aspects of the great Japanese tradition. The audience wassurprised to learn that in addition to being a top scholar and a successful professor, he is also a black-belt 7th Grand Master and a Karate instructor. In fact, he still trains martial art students once a week. He is indeed a remarkable telecom professor and well achieved personality!

IEEE Montréal ComSoc Chapter Chair, Dr. Anader Benyamin-Seeyar presenting a symbolic gift to Dr. Koichi Asatani thanking him for his lecture.

Dinner after seminar, left to right: Dr. Koichi Asatani, Dr. Anader Benyamin-Seeyar (Concordia U.), and Dr. Mouhamed Abdulla (U. Québec).

As shown in the top photo, Dr. Asatani received a token of appreciation and a memorable gift for his visit and for offering a stimulating talk. We then distributed feedback sheets to anonymously evaluate the quality of the event. It was echoed by the respondents that Prof. Asatani's talk was very interesting, informative, and highly resourceful. Following the seminar, we also had the opportunity and pleasure to entertain Dr. Asatani for a dinner where we discussed various technical, administrative, and social issues relevant to the ComSoc community. A snapshot of this moment is shown in the bottom photo.

Montréal was Dr. Asatani's first stop in a series of North American destinations. The next day, Feb. 19, he gave another talk in Québec City at Laval University. He then traveled to the St. Maurice Section and made a presentation on Feb. 20 at UQTR. On Feb. 21 he returned to Montréal and continued his journey to Austin, Texas. Dr. Asatani then delivered three lectures in the US: on Feb. 23 in Austin, on Feb. 24 in San Antonio, and finally on Feb. 25 in New Orleans. On Feb. 26 he headed back to Tokyo and safely returned home the day after.

On behalf of the IEEE Montréal Section, we would like to publicly highlight, acknowledge, and thank Prof. Koichi Asatani for his total dedication, efforts, and instrumental role within IEEE ComSoc. Also, we would like to sincerely thank Mr. Fawzi Behmann, the NA DLT/DSP Coordinator, for his hard work preparing Prof. Asatani's visit from Japan to six North American ComSoc Chapters over 10 days. The task was not simple, but he took great care to arrange the travel plans and synchronized the sequence of events with all the Chapter Chairs in a detailed and precise manner. Many thanks Mr. Behmann!

REGIONAL REPORT

Tenth Anniversary of RATEL, Serbia

By Milan Jankovic, RATEL, Serbia, and Nicolae Oaca, Romania

RATEL, the Regulatory Agency for Electronic Communications and Postal Services of the Republic of Serbia, celebrated its 10th anniversary on April 23, 2015, with the participation of the representatives of the regulatory bodies from the EU member countries, candidate countries, as well as from the Central and Eastern Regional Working Group — Bulgaria, Macedonia, Albania, Turkey, Poland, Montenegro, Slovenia, and Croatia — which shared their experiences in regulating their markets. Dr. Milan Jankovic, CEO and Director of RATEL since 2006, presented the highlights of RATEL's activities in the past 10 years. RATEL, and mainly Dr. Milan Jankovic, played a crucial role in establishing and running the Central and Eastern Regional Working Group (http://www.ceeregionalworkinggroup.net) which put together interests in our region. One of the main aims of RATEL is to put the Serbian telecommunications market in line with the EU requirements, in preparation for its admission to the EU.

Group photo of participants on RATEL's tenth anniversary.

The Current Serbian Market, and the Prime Accomplishments of RATEL

The Telecom Market by Year-End 2014: The total revenues generated by the Serbian telecom market in 2014 were €1.62 billion, or 4.5 percent of country's GDP. Mobile telephony contributed 58 percent, fixed telephony 22.4 percent, Internet 11 percent, media content distribution 8 percent, and VoIP 0.1 percent. The investments in the telecom sector amounted to €186 million. A comparative overview of the number of users and penetration rate for the public fixed communication network, public mobile communication network, Internet, and cable systems for 2012, 2013, and 2014 is given in Table 1. Among them, Internet usage has the fastest growth due to the existing high speed networks and competition.

Market Analysis and SMP Operators: Following the public consultation procedure on the report on the analysis of the wholesale market for call termination on the public telephone network, RATEL adopted a Decision on 29 December, 2014, designating SMP operators and imposing obligations. The next round of the

(Continued on Newsletter page 4)

	2012		2013		2014	
	Number (thousands)	Penetration (%)	Number (thousands)	Penetration (%)	Number (thousands)	Penetration (%)
Fixed lines	2,990.1	41.29	2,938	40.91	2,856.1	39.96
Mobile users	9,137.9	126.19	9,198.7	128.09	9,344.98	130.76
Internet subscribers	5,038.9	69.26	5,691.6	79.25	6,191,52	86.63
Cable subscribers	1,442.2	19.92	1,552.5	21.62	1,497	20.95
*Source: RATEL						

Table 1. A comparative overview of the number of users of the basic electronic communication services in the last three years.

ComSoc 2015 Election: Take Time to Vote

Ballots were e-mailed and/or postal mailed 29 May 2015 to all Higher Grade* IEEE Communications Society Members and Affiliates (excluding Students) whose memberships were effective prior to 1 May 2015.You must have an e-ballot or paper ballot before you can vote.

Vote Now using the URL below.You will need your IEEE Account username/password to access the ballot.If you do not remember your password, you may retrieve it on the voter login page.

https://eballot4.votenet.com/IEEE

If you have questions about the IEEE ComSoc voting process or would like to request a paper ballot, please contact ieee-comsocvote@ieee.org or +1 732 562 3904.

If you do not receive a ballot by 30 June, but you feel your membership was valid before 1 May 2015, you may e-mail ieee-comsocvote@ieee.org or call +1 732 562 3904 to check your member status.(Provide your member number, full name, and address.)

Please note IEEE Policy (Section 14.1) that IEEE mailing lists should not be used for "electioneering" in connection with any office within the IEEE.

Voting for this election closes 24 July 2015 at 4:00 p.m. EDT!Please vote!

*Includes Graduate Student Members

CHAPTER NEWS

Join the ComSoc Conference Chapter Challenge

Win Up To $5,000 USD to Fund Your Chapter's Activities and Functions

By Heather Ann Sweeney, IEEE ComSoc Staff

Have you joined the ComSoc Conference Chapter Challenge? There is still time to lead the way and win this year's top prize in the ComSoc Conference Chapter Challenge. As active IEEE ComSoc members, we invite you and your constituents to win up to $5,000 USD for registering in our society's ongoing flagship and core events. At the end of 2015, the three IEEE ComSoc chapters with the highest percentage of conference registrations per chapter member will be awarded valuable funds for financing anything from educational programs and networking functions to new member drives.

How to Participate: It's easy. Just do what comes naturally. Attend ComSoc flagship or core conferences, and enter your chapter code on the registration form under the "If referred by a ComSoc Chapter, please state chapter name" question. Your chapter code is your chapter name plus 2015. For example, Austria Chapter 2015 would be the code for the Austria Chapter. To increase your chapter's chances of winning, attend one or more ComSoc flagship and core conferences, and encourage your fellow chapter members to do the same!

Prizes:

1st Place: $5,000 USD
2nd Place: $3,000 USD
3rd Place: $1,500 USD

Eligibility: To be eligible to participate in the Challenge, a chapter must have at least one registration to any five ComSoc flagship and core conferences. The remaining conferences include CNS 2015, CSCN 2015, DySPAN 2015, GLOBECOM 2015, SmartGridComm 2015, CCNC 2016, or WCNC 2016. Complimentary registrations will be excluded.

Duration: The Challenge runs until December 31, 2015.

Questions: If you have any questions on this competition or would like to receive conference promotional material for distribution to your IEEE ComSoc chapter members, please contact Heather Ann Sweeney at h.sweeney@comsoc.org.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

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

IEEE COMMUNICATIONS SOCIETY

STEFANO BREGNI, VICE-PRESIDENT MEMBER RELATIONS
PEDRO AGUILERA, DIRECTOR OF LA REGION
MERRILY HARTMANN, DIRECTOR OF NA REGION
HANNA BOGUCKA, DIRECTOR OF EAME REGION
WANJIUN LIAO, DIRECTOR OF AP REGION
CURTIS SILLER, DIRECTOR OF SISTER AND RELATED SOCIETIES

REGIONAL CORRESPONDENTS WHO CONTRIBUTED TO THIS ISSUE

FAWZI BEHMANN (FAWZI.BEHMANN@GMAIL.COM)
NICOLAE OACA (NICOLAE_OACA@YAHOO.COM)
EWELL TAN, SINGAPORE (EWELL.TAN@IEEE.ORG)

A publication of the IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

TENTH ANNIVERSARY OF RATEL/*Continued from page 3*

market analysis for the remaining markets will take place in 2015.

Technological Neutrality: Since 1 January 2015, technological neutrality has been introduced in the 900 MHz and 2100 MHz frequency bands by a RATEL decision.

On 19 February, 2015 RATEL carried out the public bidding process for the issuance of individual licences for the usage of the radio frequencies in the 1710–1780/1805-1880 MHz frequency bands. Upon completion of the public bidding process, individual licences have been awarded to all three existing mobile operators, who made the payment of the one-off licence issuance fee, €7 million each, to the Treasury of the Republic of Serbia. The mobile operators provided the technology-neutral services on 25 March, 2015, so the new generation of mobile communications, 4G, is now available in the Republic of Serbia, enabling better coverage and faster Internet.

The Ministry is drafting rulebooks on minimum conditions for the issuance of the individual licence for the use of radio-frequencies in the 800 MHz frequency band and for the available 5 MHz in the 1800 MHz frequency band.

Digital Switchover: The Ministry of Trade, Tourism and Telecommunications adopted the Rulebook on the Transition from Analogue to Digital Terrestrial Television Broadcasting and Access to Multiplexes ("Official Gazette of RS," no. 86/14, 18/15 and 30/15). This Rulebook defines frequency channels for the first three multiplexes, procedure for accessing multiplexes, further development of the Initial Network (experimental networkfor the new technology), as well as the Digital Switchover Plan, which defines the timetable for transition from analogue to digital terrestrial broadcasting of the television programs per regions. The date of the analogue switch-off for the first region was 15 April, 2015, while the completion of the digital switchover was set for 17 June, 2015 in line with the international obligations of the Republic of Serbia.

The digital signal of the Initial Network broadcast by the public enterprise ETV covers more than 93 percent of the population. ETV has the obligation to cover at least 95 percent of the population with the programs in the first multiplex, and at least 90 percent for the second multiplex by June 2015.

Additional funds were allocated to the transition to digital terrestrial TV. The negotiations with EBRD were finished in September 2014 with a loan agreement of €24 million for purchasing the equipment for the distribution and broadcasting of the digital terrestrial TV signal, as well as for the reconstruction of 56 locations from which this signal will be broadcasted.

On 19 March, 2015 the Government of Serbia adopted a Decree to establish support measures and requirements for socially vulnerable consumers and the allocation of vouchers for subsidized purchase of equipment for the reception of digital television signals.

Digital Switchover Finished by 7 June, 2015: Public promotion of the digital switchover is primarily focused on the electronic media, which are obliged to support the switchover process by informing the citizens about the key issues. It also includes print media, Internet portals, and social networks, and it will also entail direct contact with citizens. This process is led by the Ministry and public service broadcaster (RTS), while the partners in the process are the Regulatory Authority of Electronic Media, RATEL, and all broadcasters (national, regional, and local).

RATEL is also an active member in the international specific entities in order to prepare national telecommunications for the EU. Due to RATEL, Serbian telecoms are now in line with the EU requirements and are waiting for admission, planned for 2020.

5G SPECTRUM: ENABLING THE FUTURE MOBILE LANDSCAPE




Hans D. Schotten *Mikko A. Uusitalo* *Jose F. Monserrat* *Olav Queseth*

The arrival of the fifth generation (5G) is expected to come together with three important enablers. First, the densification of access nodes will continue. Second, 5G networks must be highly flexible and adapt to the dynamism of the traffic location and patterns. For this, some of the radio access network (RAN) functionalities will run in large computer centers, able to dynamically assign more or fewer units of computation to the virtual cells distributed in the network. Finally, a complex landscape of spectrum availability and access will emerge where multiple frequency bands, subject to different regulations including various forms of shared spectrum, are expected to be available to wireless communication systems.

The discussions on spectrum demand in the ongoing preparation phase for the International Telecommunication Union (ITU) World Radiocommunications Conference 2015 (WRC-15) are dominated by the dramatic increase in mobile data volume caused by the rapidly growing multimedia prosumption of end users. Licensed shared access (LSA) and other concepts were developed to address the resulting challenges.

The preparation for WRC-19, and the long-term discussions on spectrum demand and usage concepts will be driven by 5G visions and, more specifically, the coherence of mobile communications solutions for many public and professional user groups within an overall mobile landscape. It is expected that this "5G" infrastructure will enable many new services in domains such as intelligent traffic systems, public safety, automation, e-health, transport, and logistics, and will help them to become economically viable. This trend will not only increase the overall spectrum demand beyond the figures discussed at WRC-15, but will result in additional requirements on spectrum resulting from the high coverage, reliability, and availability requirements of the new services to be supported.

This Feature Topic provides an overview of the major developments in the use of future 5G spectrum. The six articles included in this issue describe new approaches for the efficient use of spectrum as well as new band opportunities that will have an impact on the coming ITU-R discussions.

The first article, "Spectrum Access System for the Citizen Broadband Radio Service" by Munawwar Sohul *et al.*, presents a survey on the current discussions about LSA in the FCC framework. Special attention is paid to the concepts of dynamic frequency assignment (DFA) and interference management (IM). Requirements, capabilities, and a minimum set of information are extracted for these two functionalities.

Once the general framework is provided from the U.S. regulation point of view, the next article, "Toward Spectrum Sharing: Opportunities and Technical Enablers" by Konstantinos Chatzikokolakis *et al.*, presents the architectural framework, and the challenges and opportunities observed from the European point of view. Moreover, this article presents a fuzzy-logic-based mechanism to make the best decision concerning the type of authorization scheme. The proposed algorithm uses a number of measurements including, among others, network load, interference situation, and mobility patterns.

The third and fourth articles provide more specific mechanisms for spectrum sharing. In particular, the third article, "Coordination Protocol for Inter-Operator Spectrum Sharing in Co-Primary 5G Small Cell Networks" by Bikramjit Singh *et al.*, considers co-primary spectrum sharing among a limited number of co-located RANs belonging to different operators. In this framework, the authors propose a non-cooperative coordination protocol for mutual renting of spectrum in which operators agree on the set of negotiation rules. With low signaling overhead, no monetary transactions are involved; instead, spectrum sharing is based on a RAN-internal virtual currency. The protocol adapts to load and interference conditions, and it has proven to be efficient in small cell scenarios.

On the other hand, the fourth article, "Spectrum and License Flexibility for 5G Networks" by Adrian Kliks *et al.*, discusses the idea of flexible licensing, which provides new opportunities for spectrum holders to make additional profit by renting portions of locally unused spectrum. Sev-

eral concepts are analyzed in this framework; of special relevance is the pluralistic licensing concept, which is the focus of the article. The article also discusses the necessary regulatory decisions made globally to facilitate these new spectrum usage approaches in the context of 5G networks.

The last two articles address two specific spectrum situations, those in the United States and in China. The fifth article, "Broadcast Television Spectrum Incentive Auctions in the U.S.: Trends, Challenges, and Opportunities" by David Gómez-Barquero and Winston Caldwell, presents an overview of the future TV broadcast spectrum incentive auction in the United States, and reviews the main business, regulatory, and technical challenges for a successful auction. In this case, the United States could be the first country to make the upper portion of the 600 MHz band available for mobile broadband, and it will become one of the hot topics for WRC-19. The article also proposes a new approach for a market-driven incentive auction, in which primary users may resell frequency packs to other interested players.

Finally, the sixth article, "5G Spectrum: Is China Ready?" authored by Tan Wang *et al.*, presents the Chinese vision on 5G spectrum, including demands, potential candidate bands, and use of spectrum. Moreover, starting from the current framework of spectrum management in China, this article offers an interesting classification of services and evaluates the specific needs of bandwidth.

We would like to thank Dr. Osman Gebizlioglu, Joseph Milizzo, Charis Scoggins, and Jennifer Porcello for their continuous support and valuable comments to improve this Feature Topic. We hope that the articles herein will encourage readers of *IEEE Communications Magazine* to contribute to the discussions on the future design, development, and adoption of 5G technologies.

BIOGRAPHIES

HANS SCHOTTEN [M] (schotten@eit.uni-kl.de) is a full professor and head of the Institute for Wireless Communications and Navigation at the University of Kaiserslautern, and scientific director and member of the Management Board of the German Research Centre for Artificial Intelligence (DFKI GmbH). In 1997, he received a Ph.D. in electrical engineering from Aachen University of Technology RWTH, Germany. He has held positions as senior researcher, project manager, and head of research groups at Aachen University of Technology, Ericsson Corporate Research, and Qualcomm Corporate R&D. At Qualcomm he has been a director for technical standards and coordinator of Qualcomm's activities in European research programs. He has published over 200 technical papers, filed 15 patents, received several awards, and served as TPC co-chair of 20+ international workshops and conferences.

MIKKO A. UUSITALO [SM] (mikko.a.uusitalo@nokia. com) is manager of Radio Research at Nokia Networks. He obtained an M.Sc. (Eng) and a Dr.Tech. from Helsinki University of Technology in 1993 and 1997, and a B.Sc. (economics) from Helsinki School of Economics in 2003 (both currently known as Aalto University). He has been at Nokia since 2000 with various roles, including principal researcher and head of international cooperation at Nokia Research. He has about 95 pending or granted patents. He was elected as Chair of WWRF for 2004–2006, and is one of the founding members of WWRF as well as a WWRF Fellow. He is also a founding member of the CELTIC EUREKA initiative.

JOSE F. MONSERRAT [SM] (jomondel@iteam.upv.es) received his M.Sc. degree with high honors and Ph.D. degree in telecommunications engineering from Polytechnic University of Valencia (UPV) in 2003 and 2007, respectively. He was the recipient of the First Regional Prize of Engineering Studies in 2003 for his outstanding student record, also receiving the Best Thesis Prize from UPV in 2008. In 2009 he was awarded with the best young researcher prize of Valencia. He is currently an associate professor in the Communications Department of UPV. His current research focuses on the design of future 5G wireless systems and their performance assessment. He has been involved in several European Projects; especially significant has been his participation in NEWCOM, PROSIMOS, WINNER+, and METIS, where he is currently leading the simulation activities. He also participated in 2010 in one external evaluation group within ITU-R on the performance assessment of the candidates for the future family of standards IMT-Advanced. He co-edited the February 2011 Special Issue on IMT-Advanced Systems of *IEEE Communications Magazine* and is a co-author of *Mobile and Wireless Communications for IMT-Advanced and Beyond* (Wiley). He holds six patents and has published more than 40 journal papers.

OLAV QUESETH (olav.queseth@ericsson.com) received his M.Sc. in computer engineering in 1995 from Chalmers, Sweden, and a Ph.D. degree in 2005 from KTH, Sweden, in radio communications networks. He is currently working at Ericsson as a master researcher. He joined the 5G research project METIS in 2012 and since April 2014 has been the project coordinator. Prior to that he worked on spectrum issues in the regulatory domain in CEPT and ITU, and before that he worked with standardization of radio aspects in 3GPP. He joined Ericsson in 2007 after doing spectrum research in the Ambient Networks EU research project.

ICIN 2015: INNOVATIONS IN SERVICES, NETWORKS AND CLOUDS

CONNECTING PEOPLE, THINGS, AND MACHINES

NOEL CRESPI, INSTITUT-MINES TÉLÉCOM, TÉLÉCOM SUDPARIS, EVRY, FRANCE
EMMANUEL BERTIN, ORANGE LABS, CAEN, FRANCE

A total of 105 delegates from 26 countries representing about 60 different organizations met in Paris, France for the 16th ICIN conference to discuss new technologies from the Internet and from the telecommunications industry. ICIN 2015 took place February 17–19, with the technical co-sponsorship of the IEEE and IEEE Communications Society, and the support of Ericsson, Nokia, BroadSoft, Orange, and Institut Mines-Telecom. ICIN has a 26-year history of anticipating the key trends in communications networks and services, showcasing technologies and architectures that become vital elements of delivering services.

The acceptance rate of the conference was 36 percent for full papers. Each paper was evaluated by at least five reviewers with an average of eight reviews per paper.

The conference was structured around four tracks: Service Webification (chaired by Roch Glitho, professor at Concordia University); Network ITzation (chaired by Bruce Maggs, professor at Duke University and VP Research of Akamai Technologies); Internet of Things (chaired by Jiangtao Wen, professor at Tsinghua University); and Social Networking and Customer Relationships (chaired by Raouf Boutaba, professor at the University of Waterloo). Thirty-three research articles were presented, plus eight poster presentations.

The conference began with a tutorial focused on the Internet of Things, delivered by Roberto Minerva, head of Innovative Architectures at Telecom Italia's Future Centre. This tutorial illustrated that, in a future where almost anything will be a potential network node, market growth will be driven by the "servitization" trend (i.e. not merely selling products, but providing services surrounding products). In this context, we should anticipate a shift from a world where devices, networks, services, and information systems are clearly distinguished, to a virtual continuum between smart objects, network, processing, and storage resources.

The conference was opened by Noël Crespi (Mines-Telecom Institute), Chair of the ICIN 2015 Technical Program Committee, and Stuart Sharrock, Chair of the ICIN International Advisory Board. A series of keynote speeches were delivered around the theme for the conference. Nicolas Demassieux, VP Research from Orange, presented insights on Orange research at the dawn of the fifth 'digital' era. Alistair Urie, Director of Architecture Strategy and 5G at Alcatel-Lucent and a Bell Labs Fellow, next gave a keynote on the whys and wherefores of 5G: why, what and when 5G is coming. He highlighted the need to design energy-aware protocols and to set up APIs between applications and networks. Henning Schulzrinne, professor at Columbia University, presented a keynote on the Internet of Things – Internet in the Small. He highlighted some lessons learned from the Internet: quality is not a substitute for quantity; data link layers come and go but IP stays; the age of application-specific is over (whether for sensors, access channels, OS or protocols); and protocols matter but programmability matters much more! In this perspective, the IoT is mostly about devices and not about changing the Internet: network parts are not really new or exciting, but designing software-defined networked devices is still a challenge. Bruce Maggs, professor at Duke University and VP Research of Akamai Technologies, explained the role of Content Delivery Networks in protecting web sites from attacks, illustrating his talk with real-life attacks. This capacity is enabled by the role of CDNs in terminating TCP connections to act as service frontends. David Soldani, VP Research of Huawei, presented the pillars of Software-Defined 5G Networks: design principles, architecture, interfaces, functions, and procedures looking at the 1ms latency requirement. Jan Holler, principal researcher at Ericsson Research, detailed the usage impacts of the Internet of Things for embedding more intelligence into our networked society.

As usual, the ICIN conference is a good indicator of the trends in academic and industry research in communication networks and services. We observed that the IMS and VoLTE completely dropped off the radar; nobody seems to be betting on their capacity to provide breakthrough innovation. The main emerging topic is the webification of communication services, enabled by the WebRTC technology along with all-web systems. This is an entirely new field, and many of the usual mechanisms from communication services need to be rethought, for example signaling, identity management, media plane, and network QoS. Another avidly discussed domain was the Internet of Things. The need for interoperability was unanimously shared, e.g. with publish/subscribe mechanisms. Concerning the ITization of networks, the need for programmability by user's applications was mentioned, as well as the importance of applying up-to-date software development paradigms to SDN/NFV, e.g. DevOps, loose coupling, Hollywood principle, etc.

Additional ICIN 2015 highlights included a closing session where all the track chairs synthesized the upcoming trends in their topic. Best paper and best presentation awards were also given at the close of the conference. Paulo Chainho (Portugal Telecom), Kay Haensge and Steffen Druesedow (Deutsche Telekom), and Michael Maruschke (University of Leipzig) won the best paper award for Signalling-On-the-fly: SigOfly. Best presentation awards were given to Bichlien Hoang (IEEE Future Directions) for How Will Rebooting Computing Help IoT? and to Jiangtao Wen (University of Tsinghua) for An IoT Electric Business Model Based on the Protocol of Bitcoin. The best poster award went to Iyas Alloush, Jacques Simonin, Siegfried Rouvrais, Yvon Kermarrec (Telecom Bretagne) and Vanea Chiprianov (University of Pau) for A Model Driven Approach for Telecommunication Service Creation Environments Relying on Enterprise Architecture. Feedback on the ICIN 2015 received verbally onsite and via email after the closing ceremony was very positive.

Planning for ICIN 2016 is already underway under the leadership of Emmanuel Bertin of Orange Labs, Technical Program Chair for ICIN 2016. The conference will take place in February 2016 in Paris. For more information please visit www.icin.co.uk

Spectrum Access System for the Citizen Broadband Radio Service

Munawwar M. Sohul, Miao Yao, Taeyoung Yang, and Jeffrey H. Reed

ABSTRACT

As a part of the global effort to address the overwhelming demand for wireless broadband capacity, the wireless communities in the United States have undertaken innovative initiatives such as tiered-access to shared spectrum. The Federal Communications Commission has proposed a dynamic spectrum management framework for a Citizen Broadband Radio Service (CBRS) governed by a spectrum access system (SAS). The implementation of a SAS capable of dynamic frequency assignment and interference management is critical for the success of the proposed framework. In this paper we present the efforts toward a spectrum sharing system in the U.S. context by summarizing different interest groups' standpoint on the FCC proposed framework. We also present an example SAS architecture that accommodates the tiered access to shared spectrum and example approaches to achieve important SAS capabilities.

INTRODUCTION

Wireless communities throughout the world have recognized the shortage of spectrum for commercial broadband uses and acknowledged the urgent need for a global effort to make additional spectrum available for broadband data. The European Telecommunications Standards Institute is working toward a licensed shared access (LSA) paradigm to enable mobile broadband services in the 2.3-2.4 GHz band [1]. In the United States, following the spectrum policy of the National Broadband Plan [2] and the President's Council of Advisors on Science and Technology (PCAST) report [3], and building on the experience of television white space (TVWS) spectrum sharing [4], the Federal Communications Commission (FCC) has taken a series of steps toward sharing federal spectrum with commercial broadband applications. In 2010 the National Telecommunications and Information Administration (NTIA) identified a number of potential spectrum blocks for shared access [5]. Later that year NTIA recommended the 3550-3650 MHz (3.5 GHz) band as a "fast track" band due to its limited propagation characteristics and geographically limited incumbent operations. In 2012 the FCC issued a Notice of Proposed Rule Making (3.5 GHz NPRM) to create a new Citizens Broadband Radio Service (CBRS) centered around a spectrum access system (SAS) [6]. The 3.5 GHz NPRM encouraged deployment of small cells and introduced an innovative general authorized access (GAA) tier to facilitate opportunistic and non-interfering basis spectrum use. In 2013, responding to the comments from different interest groups, the FCC proposed an expanded eligibility for the priority access (PA) tier and issued two public notices (PNs) requesting comments on the CBRS licensing approach and SAS [1, 7]. Based on the interest groups' responses to the 3.5 GHz NPRM and licensing and SAS PNs, in 2014 the FCC issued a further NPRM with regards to the commercial operations in the 3.5 GHz band (3.5 GHz FNPRM) [8].

The proposed CBRS with its spectrum sharing concept can be a stepping stone to resolve the spectrum deficit issues and potentially evolve to 5G concepts. The SAS would take inputs from incumbents regarding their spectrum utilization and manage the secondary use of the available spectrum opportunities. But the dynamic spectrum management (DSM) approach naturally discourages investment and participation from the cellular industry by putting them outside the comfort zone of their existing technological capability and administrative structure. So the dilemma in front of the FCC is to promote DSM and at the same time ensure participation of the wireless industry in the CBRS. In this paper we will present a brief survey by summarizing the stakeholder's comments [9] in response to the 3.5 GHz FNPRM.

Our survey of this ongoing discussion yields a number of important SAS functionalities that are necessary for successful initial launch and operation of the proposed CBRS. In this paper we focus on two important SAS functionalities: dynamic frequency assignment (DFA) and interference management (IM). To address the complexity of the SAS functionalities and scale of SAS responsibilities, we believe that a modular

The authors are with Virginia Tech.

0163-6804/15/$25.00 © 2015 IEEE

Interest group		Members/commenters
SAS/TVWS administrators		Google, Telcordia Technologies, Comsearch, Spectrum Bridge
Dynamic spectrum access industry		Federated Wireless LLC, InterDigital Inc., Microsoft Corporation, Shared Spectrum Company, xG Technology Inc.
Cellular industry	Mobile network operators	AT&T, Sprint Corporation, T-Mobile, Verizon and Verizon Wireless
	Vendors	Alcatel-Lucent, Ericsson, Motorola Mobility, Nokia, QualComm
	Wireless industry associations	4G Americas, IEEE DySPAN-SC (Standard Committee), New America Foundation and Public Knowledge, Telecommunications Industry Association, PCIA — The Wireless Infrastructure Association and HetNet Forum, Wi-Fi alliance, Wireless Innovation Forum, Wireless Internet Service Providers Association
Satellite Earth station	S-Band satellite operation	Satellite Industry Association, Baron Services Inc.
	C-Band satellite operation	Content Interest, National Public Radio, National Cable and Telecommunications Association
Utilities and CII		American Petroleum Association, Entelec, Exelon, Motorola Solutions, Iberdrola USA Networks, Oncor Electric Delivery Company, Siemens Industry Inc., Southern Company Services, Utilities Telecom Council, Xcel Energy Services
Backhaul network		BLiNQ, Cohere Technologies, Sprint Corporation
Others	Terrestrial fixed microwave communications	Fixed Wireless Communications Coalition
	Communications for underserved communities and tribal lands	Blooston Coalition
	Financial organization	Cantor Telecom Services L.P.

Table 1. Interest groups among the stakeholders.

composition of SAS with close interaction among different modules will best serve the proposed access framework. Following an "F-I-C" (functionality-information-composition) approach, we try to identify the required SAS module, capabilities, and minimum set of information required for DFA and IM functionalities. The reminder of the paper is organized as follows. We present a summary of the comments for the rules that influence DFA and IM functionalities. Next, we discuss the requirements of SAS considering the inputs of the interest groups and outline a potential SAS architecture. We also introduce our initial work as example approaches to achieve two of the required SAS capabilities. Finally, we summarize the article and present suggested future work.

Overview of the 3.5 GHz FNPRM

In this section we briefly discuss only those rules that influence DFA and IM functionalities. We also present the standpoint of the major interest groups about these proposed rules. Table 1 presents a tabular representation of different interest groups along with associated organizations.

Three-Tier Access Framework

The three-tier model provides the framework for the proposed DSM approach. Existing primary operations, including federal users and grandfathered fixed-satellite service (FSS) earth stations, would make up the incumbent access (IA) tier. The CBRS would be divided into PA and GAA tiers of service, each of which would be required to operate on a non-interference basis with the IA tier. The FCC also proposed an expanded eligibility for the CBRS where the CBSDs (CBRS devices) are required to be authorized and coordinated by one or more authorized SASs.

Critics of the three-tier system, especially the cellular industry, raised two important questions: whether the proposed framework provides sufficient certainty to attract investment, and whether the existing technology is capable of supporting the proposed framework. According to the cellular industry, the primary reason for the uncertainty is the introduction of DSM concepts and the inability of the existing technology to support such a dynamic approach. Commenters outside the cellular industry mostly supported the three-tier model, mentioning that the proposed DSM approach is implementable using existing technologies. Based on the experience of the TVWS database operation, the existing geolocation database technology and spectrum management platforms are capable of facilitating tiered access to spectrum. The advocates of the DSM approach argued that the dynamic approach provides greater certainty and protection to all users, as the loss of any specific chan-

The exact partitioning of the band between PA and GAA use would be determined dynamically, based on need. In our opinion, the FCC should offer explicit incentives for a widespread acceptance of the allocation proposal as the allocation rules will define the level of dynamism in the spectrum management of the CBRS.

nel in any specific location does not result in the loss of PA license (PAL) rights due to incumbent operations. We believe that the FCC needs to work with all stakeholders to provide the necessary certainty for commercial operations.

Allocation Proposal for the CBRS

The allocation proposal for the 3.5 GHz band has four major aspects. First, it reserves 50 percent of the 3.5 GHz Band for GAA use. After accounting for IA tier use, the remaining spectrum will be assigned as PALs. Second, unassigned or unused PAL channels would be available for GAA on an opportunistic basis. Third, contained access (CA) users may request up to 20 MHz of GAA spectrum to be reserved inside a CA Facility. Fourth, the SAS dynamically assigns and maintains CBRS spectrum use in real time, and there will be no fixed spectral location for PA or GAA allocations. Although PALs would be assigned in blocks of 10 MHz, there is no fixed channel size for GAA use and the GAA users would be permitted by SAS to operate on a range of frequencies within the GAA pool.

Prospective PA users opposed both the spectrum floor and opportunistic access provisions for the GAA users. They argued that the GAA spectrum floor may not leave enough spectrum for PA users and violates the priority status of the tier. Also, if the "use-it-or-share-it" approach is adopted, the PA operator should be in charge of determining the "use" status of PALs. Outside the cellular industry, a majority of the commenters agreed with both the GAA provisions. The DSM industry requested a more aggressive reservation of a minimum of 50 percent of the available bandwidth for GAA use. The exact partitioning of the band between PA and GAA use would be determined dynamically, based on need. In our opinion, the FCC should offer explicit incentives for a widespread acceptance of the allocation proposal as the allocation rules will define the level of dynamism in the spectrum management of the CBRS.

Technical Rules for the CBSDs and General Radio Requirements

The technical rules for the CBSDs were proposed to ensure efficient spectrum use and coexistence among different tiers of CBRS. The FCC proposed that the CBSDs must be able to determine their geographic coordinates to an accuracy of ±50 meters horizontal and ±3 meters vertical. Further, a CBSD must check its position every 60 seconds and report any changes in its position within 60 seconds to the SAS. The FCC proposed to require CBSDs to be interoperable across the entire band to facilitate dynamic frequency assignment and to report any interference exceeding a threshold established by the SAS.

The general radio requirements proposed a list of conducted and emitted power limits for different operational environments (FNPRM §96.38(b) [8]). It also specified that the median signal strength of CBSD transmission on the co-channel PAL boundary shall not exceed –80 dBm/10 MHz. The rule proposed an out-of-band-emission (OOBE) limit of –13 dBm/MHz in the 3.5 GHz band, and a more stringent limit of –40 dBm/MHz with a transition gap of 30 MHz immediately outside the band. The rule also specified that the PA licensees must accept co-channel interference up to a power spectral density (PSD) level not to exceed –30 dBm/10 MHz, and the CBSDs must accept interference in authorized areas of operation from federal radar systems up to a peak field strength level of 180 dBuV/m.

The commenters acknowledged to the usefulness of geolocation reporting but pointed out the inability of the existing technology to satisfy the virtual accuracy requirement. Achieving the timing requirements may also prove to be challenging. DSM supporters suggested that SAS should be allowed to dynamically calculate worst-case interference based on the location capabilities of the device. Commenters also supported the proposed rule for interoperability that the CBSDs should be capable of dynamic frequency selection across the entire band based on an SAS channel assignment. The DSM industry supported the rule for interference reporting and agreed that SAS should use these reports to determine suitable assignments and identify sources of interference. The cellular industry proposed that the interference resolution should be left to the PA operators not to SAS.

A majority of the stakeholders requested increased power limits. If the proposed limits are adopted, SAS should have discretion to permit higher power levels based on the location of the CBSD and knowledge of incumbent operations. Rather than a one-size-fits-all threshold of –80 dBm/10MHz, commenters suggested that the PA and GAA service providers may coordinate with each other to determine an acceptable received signal strength (RSS) limit. Members of the cellular industry opposed the proposed OOBE limits as they do not comply with the 3GPP TS 36.101 limits [10]. Commenters outside the cellular industry mostly agreed with the proposed OOBE limit and opposed adoption of more stringent limits. The DSM industry suggested that appropriate OOBE limits for specific operating environments should be determined based on the need to protect specific operations at a given time, place, and frequency.

Protection of the Incumbent Operations

The 3.5 GHZ FNPRM proposed that CBSDs must comply with the proposed geographic exclusion zones set forth in the Fast Track Report to ensure compatibility with federal operations. The SAS must ensure that CBSDs do not operate within exclusion zones (EZ) and immediately suspend operation of any CBSDs found to be causing harmful interference to IA users until such harmful interference can be resolved.

Commenters from every interest group agreed that the proposed exclusion zone effectively makes the band unusable and a reduced zone is critical to the acceptance of this band. They pointed out two major drawbacks of the proposed exclusion zone. First, the analysis of the exclusion zone does not consider the opera-

tional characteristics of small cell. Second, the analysis unwarrantedly considers the protection of CBRS operation. Some of the commenters suggested that SAS should allow the CBSDs coordinated access with incumbent operations. Others suggested a more dynamic approach to the exclusion zone by adopting interference protection as a criterion based on "real world" incumbent use. In our opinion, the comments in response to the proposed exclusion zone ask for FCC initiatives to provide a more acceptable incumbent protection approach.

Perspective of the Interest Groups

The survey of the comments revealed four major interest groups outside the federal incumbents: the FSS earth stations, the utilities and critical infrastructure industry (CII), the cellular industry, and the advocates of the DSM approach. The non-federal incumbents are concerned about ensuring comfortable interference environments for their operations. As expected, the strongest opposition against the DSM approach came from the cellular industry. The survey also indicated that the cellular industry would find them in a more favorable position if some of their more justified concerns, such as determination of the PAL "use" status, resolution of any interference scenario, and end-user frequency assignment, are addressed. Although the DSM supporters argue for a more aggressive role of the dynamic SAS to manage spectrum usage, the FCC can accommodate the above mention concerns of the cellular industry without compromising its vision of dynamic spectrum management. To resolve any unwanted interference issue, the SAS can instruct with the PA operators to adjust their operating parameters. Also, SAS can avoid micro-management by dynamically allocating the PALs to the PA service providers and leaving end-use frequency assignment to the PA operator. The SAS can also include the PA user's input to determine the PAL "use" status. These simple modifications in the proposed rules will allow the PA operators more control over their operations, and at the same time SAS will be able to manage the dynamic spectrum sharing system more efficiently.

SAS Architecture for DFA and IM Functionalities

In this section we will discuss the requirements of the SAS to carry out the DFA and IM functionalities considering the inputs of the interest groups and outline a potential SAS architecture. In this example SAS architecture we trust the PA operator to perform the end-user frequency assignment and interference resolution of the PA tier. The SAS would include the PA operator's inputs while determining the PAL "use status".

SAS Requirements

In order to efficiently execute the functionalities, SAS must be able to coordinate with the incumbent and CBRS operations. Both DFA and IM functionalities will require the SAS to access the incumbent database to gather information about incumbent spectrum usage. Depending on the information provided by the incumbent, the SAS may need to have the capability of using a geolocation database and/or detecting the incumbent operation through sensing. SAS also needs to coordinate with the CBRS operations such as authorizing the CBSDs, gathering CBSD information including geolocation, interference environment, radio parameters, and spectrum demand with associated QoS requirements. SAS should be able to dynamically allocate spectrum to different tiers of the CBRS, and detect and resolve any unwanted interference event.

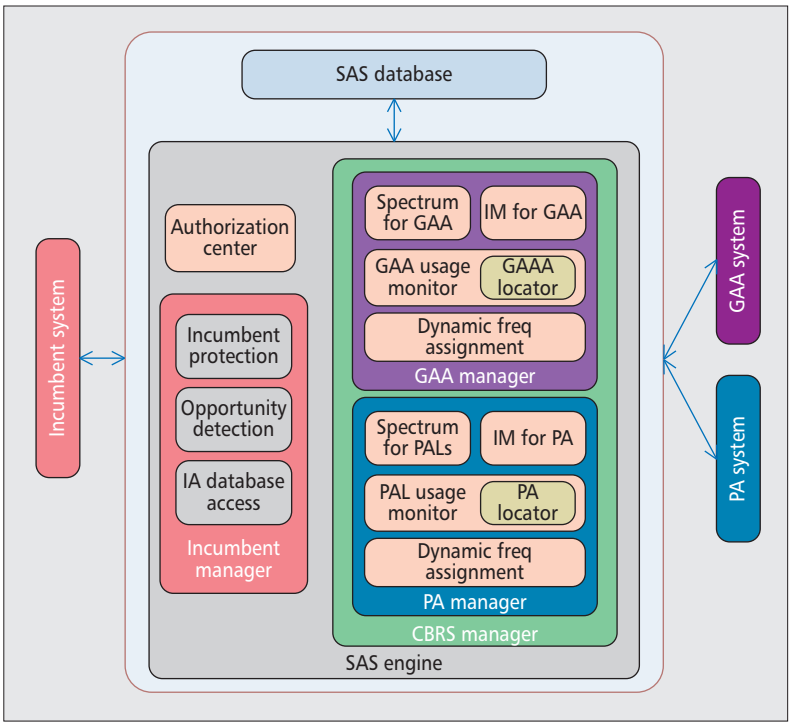


Figure 1. SAS architecture for DFA and IM functionalities.

Example SAS Architecture

The example SAS architecture has four major modules: the incumbent manager, the CBRS manager, the authorization module, and the SAS database (Fig. 1). The CBRS manager has two separate modules, the PA manager and the GAA manager, to facilitate the operation of the two tiers in the CBRS.

Incumbent Manager: The responsibility of the incumbent manager is to ensure coordination with the incumbent operations. The database access module fetches information based on the prior agreement with the incumbent. The incumbent may agree to provide its usage information such as duration of operation or only the operational parameters such as transmit power and location, antenna height, and protection contour. The spectrum detection module identifies the available opportunities based on the information gathered by the database access module. If the SAS is provided with the incumbent's operational parameters, it can either use the geolocation database, sensing technologies, or a database-plus-sensing approach to identify the available spectrum

Type of information		Information set
CBSD authorization information		• Requested authorization status • CBSD location • Area of operation • Antenna parameters • FCC identification information • Contact information
Spectrum opportunity information	Assurance from the incumbent	• List of available opportunities • Duration of opportunity • Probability of PU return
	Geolocation database approach	• Incumbent information –Transmitter identity and geolocation –Protection contour –Transmission power –Antenna parameters –Interference tolerance capability • Terrain profile • Propagation model
	Incumbent detection through sensing	• Sensing node location • Sensing node reputation • Fusion center location
Spectrum information for CBRS	PA tier	• Number of PAL channels available • Number of active PAL channels
	GAA tiers	• Spectrum allocation for CA users • Spectrum available for GAA users • PAL channels used by GAA users
Frequency assignment information		• Assigned PALs • Assigned GAA frequencies • Unused PA frequencies assigned for GAA use • Assigned frequencies to CA users • Available spectrum for future assignment
CBRS interference environment		• Interference environment among neighboring PALs • Interference from GAA to PALs • Interference environment of PALs near IA protection zones • Interference environment of GAA near IA protection zones • Interference environment of PALs near CAF
Incumbent interference environment	Interference environment of the incumbent	• Protection zones for the incumbent • Coordinated access agreements • CBRS frequency assignment near protection zones • Incumbent interference tolerant capability • Unwanted interference event –Geographic area of interference –Frequency of interference –Time to vacate for CBRS –Suggested radio parameters for CBRS
	CBRS usage information	• List of CBSDs near incumbent protection zone • Interference environment of PALs near the protection zone • Interference environment of the GAA near the protection zone

Table 2. SAS database information set.

opportunities. If any interference event is identified, the database access module can gather information about the spectral and geographical location of the source and acceptable "time to vacate" parameter. The incumbent protection module should have the ability to inform the IM module of the PA and GAA managers if the incumbent suffers from any unwanted interference.

Authorization Center: The authorization module helps the SAS to effectively carry out both the DFA and IM functionalities. This module should have the ability to receive information from the CBRS users, perform the authorization procedure, and inform the applica-

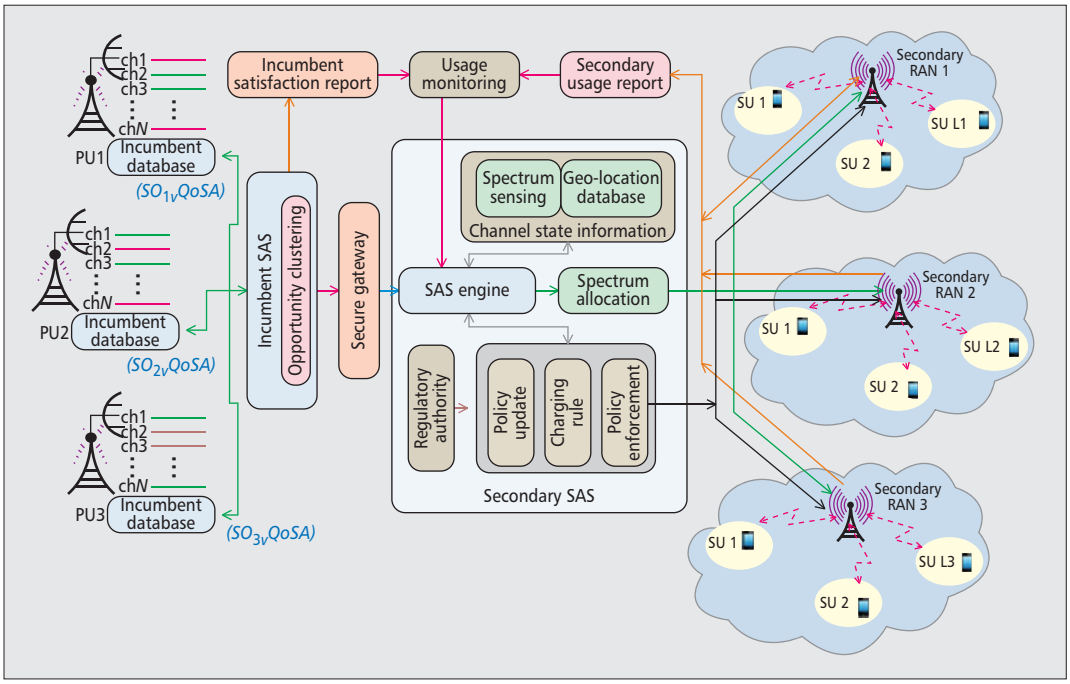


Figure 2. QoS assurance for shared spectrum systems [11].

The SAS database gathers information on the CBSDs' operational parameters, interference environments, and spectrum usage and availability information for each of the CBRS tiers. Also, the SAS database needs to gather information from the incumbent database about its spectrum usage, operational parameters, and interference environment.

tions about their authorization status. Information provided by the CBSDs for authorization purposes allows the SAS to monitor the CBSDs' operations and ensure that they are abiding by the interference protection requirements of the incumbents. The PA and GAA manager modules use this information to effectively execute the DFA and IM functionalities.

PA Manager: The PA manager ensures the available opportunities and dynamically allocates spectrum for PA operation. This module needs to coordinate with both the IA and GAA manager to obtain the spectrum availability information. The spectrum availability module prepares a list of PAL channels available for future assignment and also informs the GAA manager if any PAL channel assigned to GAA users is to be claimed back. The usage monitor module is responsible for determining the usage status of the PAL using the geolocation reports from the PA users. The DFA module ensures dynamic assignment of available PA spectrum with a best effort approach toward contiguous allocation. The IM module ensures that the users are abiding by all the proposed radio requirements such as transmission power and OOBE limits and maintains a comfortable interference environment among the neighboring PA users.

GAA Manager: The GAA spectrum availability module ensures the availability of the GAA spectrum pool and unused PAL channels, and accounts for the reserved CA spectrum. This module coordinates with the spectrum availability modules of the IA and PA managers and generates a list of available frequencies for the GAA users. The DFA module receives the spectrum demands from the GAA users with associated QoS requirements and dynamically schedules the available opportunities. The GAA usage monitor and IM modules facilitate an acceptable interference environment such that the GAA devices do not cause any harmful interference to the higher tier users. If any unwanted interference event is identified, the IM module advices the GAA device to adjust its operating parameters or informs the enforcement entity to ensure a desired interference environment.

SAS Database: The SAS database gathers all the information required by the above mentioned modules to perform the DFA and IM functionalities (Table 2). The set of information presented here serves as an initial framework for the SAS database. The SAS database gathers information on the CBSDs' operational parameters, interference environments, and spectrum usage and availability information for each of the CBRS tiers. Also, the SAS database needs to gather information from the incumbent database about its spectrum usage, operational parameters, and interference environment.

Example Approaches for SAS Capabilities

SAS requires a number of capabilities to effectively perform the DFA and IM functionalities. For example, effective duration of a spectrum opportunity is an important factor for delivering any desired level of QoS to the secondary users (SUs) in shared bands. Also, the use of dedicated sensors deployed in key locations (e.g. along the coast) could significantly improve opportunity detection and incumbent protection. In this section we present two example approaches that could help SAS to carry out the DFA and IM functionalities in a more efficient way.

Duration of Spectrum Opportunity

Our work in [11] proposed a QoS assurance (QoSA) approach, in place of QoS prediction, centered around the SAS to get an idea about

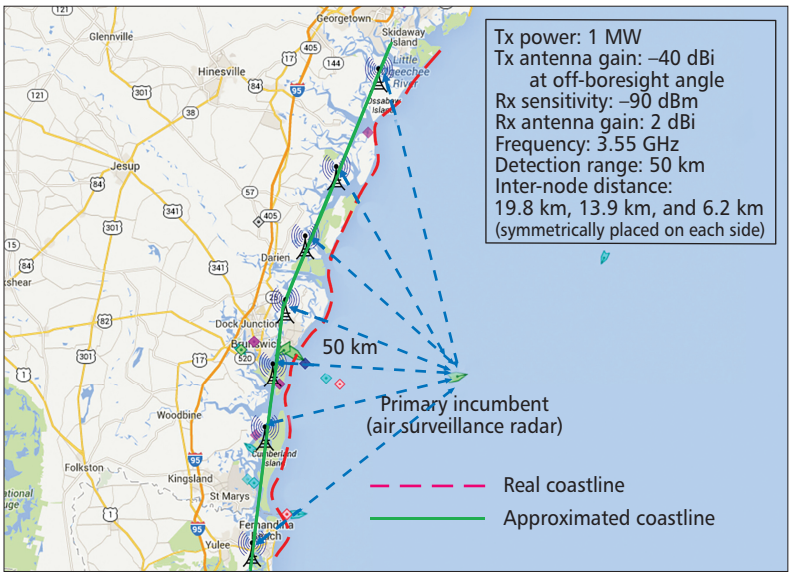


Figure 3. Illustration of sensor placement for incumbent detection and simulation context.

the opportunity duration. QoSA is a probabilistic assurance about each of the idle primary users' (PU) channels stating that the channel under consideration will be available for secondary use for certain duration of time with probability QoSA. The QoSA approach requires minimal involvement from the PU and reduces the likelihood of exposing sensitive PU information using an "opportunity clustering" mechanism.

The SAS framework proposed to incorporate the QoSA approach has two major blocks: the incumbent SAS (I-SAS) and the secondary SAS (S-SAS) (Fig. 2). I-SAS gathers the usage statistics and the tentative future usage plan from the incumbent database and calculates the probabilistic assurance for each of the idle channels. I-SAS then generates clusters of similar quality opportunities based on the associated QoSA and forwards it to the S-SAS through a secured gateway. Based on QoSA, the S-SAS helps the SU resource manager (SU-RM), serving multiple SU with different QoS demands, to perform an assignment problem that maximizes the overall spectrum efficiency for the shared bands.

In the existing approaches, SU is responsible for providing interference protection to the PU, whereas the QoSA approach allows the PU to influence SU operation through SAS. The PU can place additional guard time and guard band to improve its interference protection. As the SU-RM knows the probabilistic assurance about the duration of the opportunity, it can plan ahead to vacate the channel and avoid interference.

The QoSA approach improves the overall SU throughput by reducing the SU's dependency on the QoS prediction, protecting the SU from the PU's interference, and reducing probability of missed opportunities. Also, it provides the SU with an indication of the achievable QoS and improves the opportunity scheduling. The simulation results of [11] showed that the QoSA approach improves the SU performance with minimal increase in the collision time.

Detection of Incumbent through Sensing Nodes

Incumbent detection using dedicated sensor nodes deployed in key locations has been suggested as one of the mechanisms to identify spectrum opportunities. While detecting the incumbents, use of multiple sensor nodes is helpful to deal with the issues such as multipath and shadowing. Although the correlation between different sensor nodes depends on the inter-node distance, the availability of multiple sensor nodes still increases the spatial diversity and reduces the probability of deep fading across all the sensor nodes. For sensor nodes located in the predetermined area with uniform distribution and within the interference range of the PU, simulation results in [12] showed that the probability of detection improves with the number of sensor nodes. In the context of maritime incumbent detection, we tried to find the sensor locations along the coastline to minimize the total number of sensors while achieving the desired probability of detection.

We used a greedy algorithm based sensor placement similar to the constant-factor approximation guarantee algorithm applied in [13]. The selection of the set of sensors was initiated with the first sensor at the closest point in the line to the incumbent target, while the greedy algorithm keep adding new sensors, which satisfies the tradeoff between correlation and path loss. During each step of the greedy algorithm to search the next node, we only consider the correlation between consecutive nodes as the correlation from other nodes can be considered negligible. Simulation results suggested that a detection probability of 0.99 with a false alarm rate of 0.05 can be achieved with as little as seven sensing nodes (Fig. 3). Although this result simplified the sensing problem, it implies that incumbent detection with dedicated sensor nodes can be a practical solution to consider.

Conclusion

To address the demand for wireless broadband capacity, the wireless communities throughout the world have taken significant steps including creative solutions such as spectrum sharing. In the United States, the FCC proposed to create a Citizen Broadband Radio Service centered around a spectrum access system (SAS) to share the 3.5 GHz band with the incumbents. Critics of the proposed framework are doubtful whether the proposed framework will be able to provide sufficient certainty to ensure the return of investment. According to these critics, the development of a SAS with functionalities of such scale and complexity will take time and delay the shared use of the band. For widespread acceptance of the approach, SAS must be able to effectively carry out two important functionalities: dynamic frequency assignment and interference management. In this paper we presented a summary of different interest groups' standpoint on the proposed rules that influence the DFA and IM function-

alities and attempted to find the required SAS composition and capabilities. This paper discussed the efforts toward a spectrum shared system in the U.S. context and looks forward to initiating a more focused analysis of the required SAS composition and functionalities to support the dynamic management of the 3.5 GHz band. The successful implementation of the dynamic spectrum management approach will significantly improve the spectrum usage efficiency and influence the management approach of other spectrum bands. This calls for a coordinated effort from government, industry, and academia to expedite the development of the dynamic SAS and move forward toward the dynamic spectrum management regime.

References

[1] T.E.T.S. Institute, "ETSI TR Mobile Broadband Services in the 2300–2400 MHz Frequency Band under Licensed Shared Access Regime," vol. V1.1.1, ed. France: ETSI, 2013.

[2] FCC, "National Broadband Plan: Connecting America," retrieved September vol. 14, ed, 2010, p. 2010.

[3] "Report to the President: Realizing the Full Potential of Government-Held Spectrum to Spur Economic Growth," PCAST, Ed., ed. White House, Washington, 2012.

[4] FCC, "Notice of Proposed Rulemaking, In the Matter of Unlicensed Operation in the TV Broadcast Bands (ET Docket No. 04-186) and Additional Spectrum for Unlicensed Devices Below 900 MHz and in the 3 GHz Band (ET Docket No. 02-380), FCC 04-113," ed: May, 2004.

[5] NTIA, "Assessment of the Near-Term Viability of Accommodating Wireless Broadband Systems in the 1675–1710 MHz, 1755–1780 MHz, 3500–3650 MHz, 4200–4220 MHz, and 4380–4400 MHz Bands," ed: Fast Track Evaluation) (Oct. 2010).

[6] FCC, "FCC NPRM In the Matter of Amendment of the Commission's Rules with Regard to Commercial Operations in the 3550–3650 MHz Band," FCC, Ed., ed. Washington, D.C. 20554, 2012.

[7] FCC, "Commission Seeks Comment on Licensing Models and Technical Requirements in the 3550-3650 MHz Band," vol. Docket 12-354, ed. Washington D.C. 20554, 2013.

[8] FCC, "FNPRM: Ammendment of the Commission's Rules with Regard to Commercial Operations in the 3550-3650 MHz Band," vol. Docket 12-354, ed. Washington D.C., 2014.

[9] Comments in Reponse to the FCC 3.5GHz FNPRM available: http://apps.fcc.gov/ecfs/ (Proceeding No 12-354).

[10] E.U.T.R. Access, "User Equipment (UE) Radio Transmission and Reception, 3GPP Std. TS 36.101," ed.

[11] M. M. Sohul *et al.*, "Quality of Service Assurance for Shared Spectrum Systems," *IEEE Military Commun. Conf. (MILCOM)*, 2014 , 2014, pp. 1471–76.

[12] W. C. Headley, V. G. Chavali, and C. R. C. M. da Silva, "Exploiting Radio Correlation and Reliability Information in Collaborative Spectrum Sensing," *IEEE Commun. Lett.*, vol. 15, 2011, pp. 825–27.

[13] C. Guestrin, A. Krause, and A. P. Singh, "Near-optimal sensor placements in gaussian processes," *Proc. 22nd Int'l. Conf. Machine learning*, 2005, pp. 265–72.

Biographies

Munawwar M. Sohul (mmsohul@vt.edu) received the M.Sc. degree in communication and signal processing from Imperial College, London, U.K., and the B.Sc. degree in electrical and electronics engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh. Currently he is working toward the Ph.D. degree in electrical engineering in the Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic and State University, Blacksburg, VA, USA, working with Dr. J. H. Reed. His research interest in wireless communication and signal processing includes spectrum sharing, opportunistic spectrum access, carrier aggregation, self-organizing networks, and cognitive radios.

Miao Yao (miaoyao@vt.edu) received a B.S. degree in electrical engineering from Beijing Jiaotong University in 2007 and an M.S. degree in microelectronic and solid state electronics from Tsinghua University in 2010. He is currently working toward his Ph.D. degree at Virginia Tech. His primary research interests are dynamic spectrum access related research issues and energy efficient resource allocation algorithms in next generation cellular communication systems.

Taeyoung Yang (mindlink@vt.edu) received the M.S. and Ph.D. degrees in electrical engineering from Virginia Tech in 2003 and 2012, respectively. He received B.S. and M.S. degrees in electronic engineering from Sung-Kyun-Kwan University, South Korea, in 1997 and 1999, respectively. He has strong theoretical background in radiation physics, co-site interference, and size-performance limits of wireless devices. Along with the theoretical background, he also has extensive hands-on experience in antennas, RF, sensors, and measurements. While he was one of the core faculty members at Wireless @ Virginia Tech in 2013 and 2014, he contributed to the development of core technologies for opportunistic spectrum access and cognitive wireless infrastructure. He is currently a research scientist at Intel Corporation.

Jeffrey H. Reed [F] (reedjh@vt.edu) is the Willis G. Worcester Professor in the Bradley Department of Electrical and Computer Engineering at Virginia Tech. He currently serves as Founder of Wireless @ Virginia Tech, one of the largest and most comprehensive university wireless research groups in the U.S., and is the Founding Faculty Member of the Ted and Karyn Hume Center for National Security and Technology. Since joining Virginia Tech in 1992 he has been PI or co-PI of approximately 100 projects covering areas such as software radio, cognitive radio, ultra wideband, and channel modeling. He is cofounder of CRT Wireless, a company that is developing cognitive radio techniques for commercial and military systems, and Power Fingerprinting, a company focused on embedded device security. He has served on panels, coordinated numerous workshops and conferences, and served on advisory groups for the Department of Commerce, Department of Defense, the State of Virginia, and NSF, as well as technical advisory boards for many companies. Recently he served as associate editor for the *Proceedings of the IEEE* issues on cognitive radio. He is a Fellow of the IEEE for contributions to software radio and communications signal processing and for leadership in engineering education, and is a past recipient of the College of Engineering Award for Excellence in Research. He is the author of three books and more than 200 journal and conference papers. Dr. Reed has two new books scheduled for publication in 2012 in the areas of cellular communications and software defined and cognitive radio. He received his B.S., M.S., and Ph.D. degrees from the University of California, Davis.

The successful implementation of the dynamic spectrum management approach will significantly improve the spectrum usage efficiency and influence the management approach of other spectrum bands. This calls for a coordinated effort from government, industry, and academia.

Toward Spectrum Sharing: Opportunities and Technical Enablers

Konstantinos Chatzikokolakis, Panagiotis Spapis, Alexandros Kaloxylos, and Nancy Alonistioti

ABSTRACT

The vast increase in the number of mobile devices and their mobile traffic demands indicates the need for additional spectrum for cellular communications. Since it is not trivial to allocate exclusively new spectrum bands for cellular communications, it is imperative to improve the spectrum usage through new spectrum sharing mechanisms. This implies that the mobile network operators will have to cooperate and interact to cover the augmented traffic requirements. In this article we present a novel architectural framework that enables the mobile network operators and other spectrum license holders to exchange information about spectrum availability. We also present a novel spectrum sharing mechanism based on fuzzy logic to facilitate operators in selecting the most suitable spectrum to cover their needs.

INTRODUCTION

Traffic analysis indicates that mobile and wireless networks will have to cope with a huge increase of data traffic over the next decade. This will occur due to the vast proliferation of mobile and wireless devices and the growing need for traffic volume per subscriber. More specifically, over 4 billion mobile devices (i.e., laptops, tablets, and smartphones) exist in the mobile service market, and an 11-fold increase of mobile data traffic by 2018 compared to 2013 is expected [1]. In addition, the actual traffic volume per subscriber increases 25–40 percent per year exceeding; hence, the expectations set by the International Telecommunication Union (ITU) [2, 3].

The Third Generation Partnership Project (3GPP), motivated by the increased mobile data traffic volume, has encouraged the research community to move in three directions:

- Spectral efficiency improvement
- Higher network cell density
- Exploitation of underutilized radio spectrum resources [4]

The first solution includes coordinated multipoint (CoMP) transmission using sophisticated multiple-input multiple-output (MIMO) techniques and interference management mechanisms. The second area deals with the addition of extra layer cells in the network with base stations (BSs) that cover smaller areas compared to macro and micro BSs. These solutions include femtocells and the use of relay nodes.

In this article, we focus on the third aspect, which deals with the extension of spectrum opportunities for mobile broadband access. Nowadays, a mobile network operator (MNO) acquires from the national regulatory authority (NRA) spectrum resources, which through network planning are allocated in different geographical areas. The reallocation of underutilized spectrum resources to congested areas is a slow process that requires thorough investigation of underused or overloaded frequency bands (for months or even years) and may require spectrum refarming. However, in the near future spectrum scarcity will call for more dynamic, flexible, and fast solutions. Up to now, flexible radio spectrum management has not been a preferable option because the allocated spectrum is more or less sufficient to cover the current needs of MNOs. Moreover, MNOs are reluctant to share or acquire spectrum to/from other operators or licensed spectrum users. In the near future, spectrum sharing will be a necessity since the aforementioned alternative solutions (e.g., MIMO, CoMP) will not be able to cover MNOs' needs in capacity. Thus, new schemes to identify spectrum opportunities and decide the most profitable choice for MNOs have to be designed.

Toward efficient spectrum management, several authorization regimes and spectrum access schemes are currently used or being investigated. These regimes may be divided into two categories based on the license that is provided to the users: individual authorization approaches and general authorization approaches (also known as license exempt or unlicensed).

In individual authorization regimes we distinguish exclusive, co-primary and licensed shared access (LSA) schemes. In general authorization regime, unlicensed shared access (i.e. Wi-Fi, Bluetooth, etc.), and secondary horizontal shared access (i.e. TV White Spaces) are used, but do not manage to fill the capacity gap due to lack of quality of service (QoS) guarantees for end users

Konstantinos Chatzikokolakis and Nancy Alonistioti are with National and Kapodistrian University of Athens.

Panagiotis Spapis and Alexandros Kaloxylos are with Huawei Technologies.

0163-6804/15/$25.00 © 2015 IEEE

[5]. On the other hand, co-primary and LSA authorization options have recently been introduced as complementary methods to the exclusive use so as to ensure predictable QoS for the users. Co-primary authorization assumes spectrum trading among MNOs in their exclusive spectrum. The LSA authorization option enables the MNOs to use spectrum allocated to an incumbent user.[1] The fact that the co-primary and LSA authorization options provide predictable QoS to end users is the rationale behind their popularity in the research community and in standardization groups. However, it should be noted that until now there has been no unified approach that will enable MNOs to advertise/discover and acquire spectrum chunks in a specific geographical area. In this article we introduce an architectural framework that addresses this issue for the co-primary and LSA spectrum cases.

The previously described spectrum flexibility, which could potentially solve the problem of spectrum scarcity, has also raised questions regarding the way an MNO could identify the most suitable spectrum fraction for covering its needs. Different approaches have been used, such as game theory [6] and interference mitigation [7]. However, these solutions identify the need for additional spectrum chunks and proceed in spectrum acquisition or bids for spectrum, without considering the special characteristics of the traffic or its sources (e.g., the mobility of users). This implies that an MNO may proceed in acquiring spectrum not suitable for the traffic characteristics and the mobile users it wants to support. Since this is a complex problem that requires the evaluation of multiple criteria we introduce a fuzzy-logic based mechanism that can be used by an MNO to decide which type of authorization scheme to select based on different monitoring measurements such as network load, interference, and mobility of users.

The rest of the article is organized as follows. We present the multi-operator spectrum sharing schemes. We illustrate a new architecture that combines LSA and co-primary spectrum sharing information to assist MNOs in discovering spectrum opportunities. We also describe in detail the information that needs to be exchanged for such an interaction. We propose a scheme based on fuzzy logic for identifying the proper authorization option for handling the user's (and the MNO's) requirements in a specific geographical area. We conclude the article and discuss the outcomes of our research.

Multioperator Spectrum Sharing Schemes

Co-Primary Sharing

Co-primary sharing enables joint use of parts of licensed spectrum between primary license holders with the same regulatory status (e.g., MNOs) [8]. Some schemes have been proposed in the literature focusing on dynamic spectrum access in multi-operator environments. The decisions on spectrum sharing among MNOs may be based on solving optimization problems (e.g., modeled as competitive game theoretic approaches [9]). Other schemes may rely on the cooperative approaches (e.g., existence of common spectrum pool after mutual agreement [10]). However, such solutions focus on the identification of spectrum needs without delving into detail regarding the information exchange for realizing the sharing concept. Reference [11] provides an architectural framework focused on sharing common spectrum resources. The proposed framework is driven by three scenarios based on the timescale of the sharing, that is, short, medium, and long term. In the short-term scenarios, a mobile user will cover her needs using spectrum chunks from another operator (i.e., not the one to which she is linked with a contract) thus requiring inter-operator agreements. In the medium-term scenarios the operator will proceed in negotiations with another operator and eventually buy a set of spectrum chunks for a short to medium timescale. Finally, in the long-term scenarios the spectrum chunks are "rented" for minutes or even hours between MNOs, thus enabling the spectrum licensee to proceed in long-term strategy planning. In medium- and long-term scenarios a mechanism for coordinating the overall procedure is required for charging and avoiding use of the same spectrum chunks simultaneously from two or more entities. However, the proposed scenarios in [11] do not focus on information exchange or provide a holistic architectural view of the network. In this article we focus on providing a common framework for the last two cases where two or more operators need to cooperate in order to buy/sell chunks of spectrum in an area. In addition, our proposed framework is also extended to other sharing schemes such as the LSA.

Licensed Shared Access

The LSA concept is a complementary regulatory paradigm to other schemes (i.e., exclusive access, unlicensed shared access, etc.) that enables sharing of spectrum between the initial rightful users of the frequency band (e.g., radio telescope) and a limited number of new users (e.g., MNOs), which will be granted individual authorizations and will be able to use the spectrum according to the sharing rules included in their rights of use [12]. The former will be referred to as incumbent users for the rest of this document, and the latter ones will be referenced as LSA licensees. Although an LSA licensee could be anyone interested in acquiring additional spectrum, we focus on MNOs as the growth of mobile data traffic and the spectrum scarcity problems indicate that this will be the first practical use case of LSA to provide access to additional spectrum [13].

In the process of spectrum sharing using the LSA concept, spectrum usage rights will be granted to LSA licensees by the NRA who is responsible for protecting the incumbent users. Furthermore, access exclusivity to spectrum resources over time, space, and frequency domains among LSA licensees is required in order to avoid mutual interference among them and deliver services with predictable QoS. However, our view is that an LSA system should provide the flexibility of spectrum sharing between two or more LSA licensees as such mutual agreements should not be excluded. More specifically, spectrum acquired by an LSA licensee should be

Toward efficient spectrum management, several authorization regimes and spectrum access schemes are currently used or being investigated. These regimes may be divided into two categories based on the license that is provided to the users: individual authorization approaches and general authorization approaches.

[1] Incumbent users in LSA are considered any license holders (e.g., radio telescopes, U.S. Coast Guard) except MNOs.

Our view is that an LSA system should provide the flexibility of spectrum sharing between two or more LSA licensees as such mutual agreements should not be excluded. More specifically, spectrum acquired by an LSA licensee should be exploitable in several ways from exclusive use to mutual renting for the specific time period for which it is given to the licensee.

exploitable in several ways from exclusive use (currently covered by the RSPG and ECC) to mutual renting for the specific time period for which it is given to the licensee. However, our view implies that the incumbent user may not be aware of the actual user of the spectrum, which shall be tackled by the NRA.

The LSA scheme is an alternative to the fixed allocation of frequency bands that has many advantages for the incumbents and licensees compared to other spectrum sharing schemes. It offers the incumbent user the confidence that his use of spectrum will remain unhindered, and LSA licensees are authorized to exploit unused spectrum provided by incumbents, which implies that they will enjoy high QoS without harmful interference from the incumbent. However, it should be noted that there is no obligation in the LSA system to coordinate spectrum access of multiple LSA licensees, which may lead to interference among them.

Combined Model for LSA and Co-Primary Sharing

Based on the previous discussion it is clear that an MNO that wishes to discover spectrum opportunities will have to select between more than one spectrum sharing schemes. In this section, we present a unified architectural framework that enables MNOs to easily discover available spectrum chunks under LSA or co-primary spectrum sharing schemes for a specific geographical area. We also discuss in detail the required information to discover any spectrum availability.

Combined Functional Architecture for LSA and Co-Primary Sharing: Standardization bodies and NRAs have been working toward ensuring all the assumptions, requirements, and obligations of the involved actors in spectrum sharing schemes. For the description and analysis of LSA schemes, the Conference of Postal and Telecommunications Administrations (CEPT) has set up a Working Group [13] to prepare a report on an LSA regulatory framework in collaboration with European Telecommunications Standards Institute (ETSI) Reconfigurable Radio Systems (RRS) Working Group 1 (WG1). The former aimed to provide guidelines to CEPT administrations in relation to the implementation of LSA and establish the regulatory provisions for LSA implementation in the 2.3 GHz band. The latter has recently published a System Reference document highlighting the applicability of mobile broadband services in the 2.3–2.4 GHz frequency band under an LSA regime [14]. Until now, this process has produced a high-level functional view. However, the analysis is still in a very early stage and does not describe in detail eithr the functionalities of the LSA system or the information that needs to be exchanged. On the other hand, co-primary spectrum sharing has yet to be defined with regard to the involved functional entities and the communication among the actors. In this article, we propose a novel and enhanced functional architecture that unifies the model of LSA and co-primary sharing. Note that although in the proposed architecture the presence of a central logical entity for coordinating the MNOs is assumed, such an entity could actually be distributed to more than one physical entity.

The proposed architecture, as illustrated in Fig. 1, is based on the architecture components introduced by CEPT for LSA, and is enhanced with coordination and spectrum controller components to extend the applicability of the model to a co-primary sharing scheme. The fundamental entities of the proposed architecture are the following:

NRA/regulator: This guarantees and provides use of rights to the actor requesting spectrum resources (for LSA and co-primary spectrum sharing) and will determine LSA spectrum award rules for the incumbent in the LSA case.

Incumbent User: This is only related to the LSA sharing scheme, and its role is to provide information for the unused spectrum to the LSA repository, which will be used by LSA licensee(s).

MNO: It is granted authorization access from the NRA and exploits the unused spectrum offered by other actors (incumbent users in LSA, other MNOs in co-primary sharing).

LSA repository: This maintains the input/data and the technical and security requirements received from the incumbent users.

LSA information translator: It constructs a model that indicates the way spectrum is consumed by the incumbent users in terms of time, space, and frequency, based on the aggregated data stored in the LSA repository. Then this information is translated to its complementary data, which indicate spectrum availability. The produced information is structured in GeoSpectrum schedules. A GeoSpectrum schedule is a construct that includes location information and spectrum usage information. Further details are given in the next subsection.

Spectrum pool: This aggregates information related to available spectrum, either offered by an incumbent user or mutually agreed among MNOs.

Coordination entity: It is responsible for handling requests for spectrum from MNOs and providing them with spectrum resources. The coordination entity is also responsible for querying the spectrum pool based on the spectrum requests from MNOs.

Spectrum controller: This is part of the MNO and comprises two subelements: the LSA controller and the co-primary controller. The spectrum controller is responsible for sending requests for additional spectrum resources to the coordination entity. A decision making process that takes into consideration various network monitoring measurements (traffic load, packet error rate, latency, etc.) is included in this entity to determine whether LSA or co-primary resources are more suitable for handling the increasing network demands. Furthermore, it is responsible for providing the spectrum availability information received from the coordination entity to the operations, administration, and maintenance (OAM) block of the MNO.

OAM: In case the network lacks resources, the OAM entity will detect the corresponding event and notify the spectrum controller. If a new spectrum chunk is acquired, the OAM will allocate the new resources to the network elements.

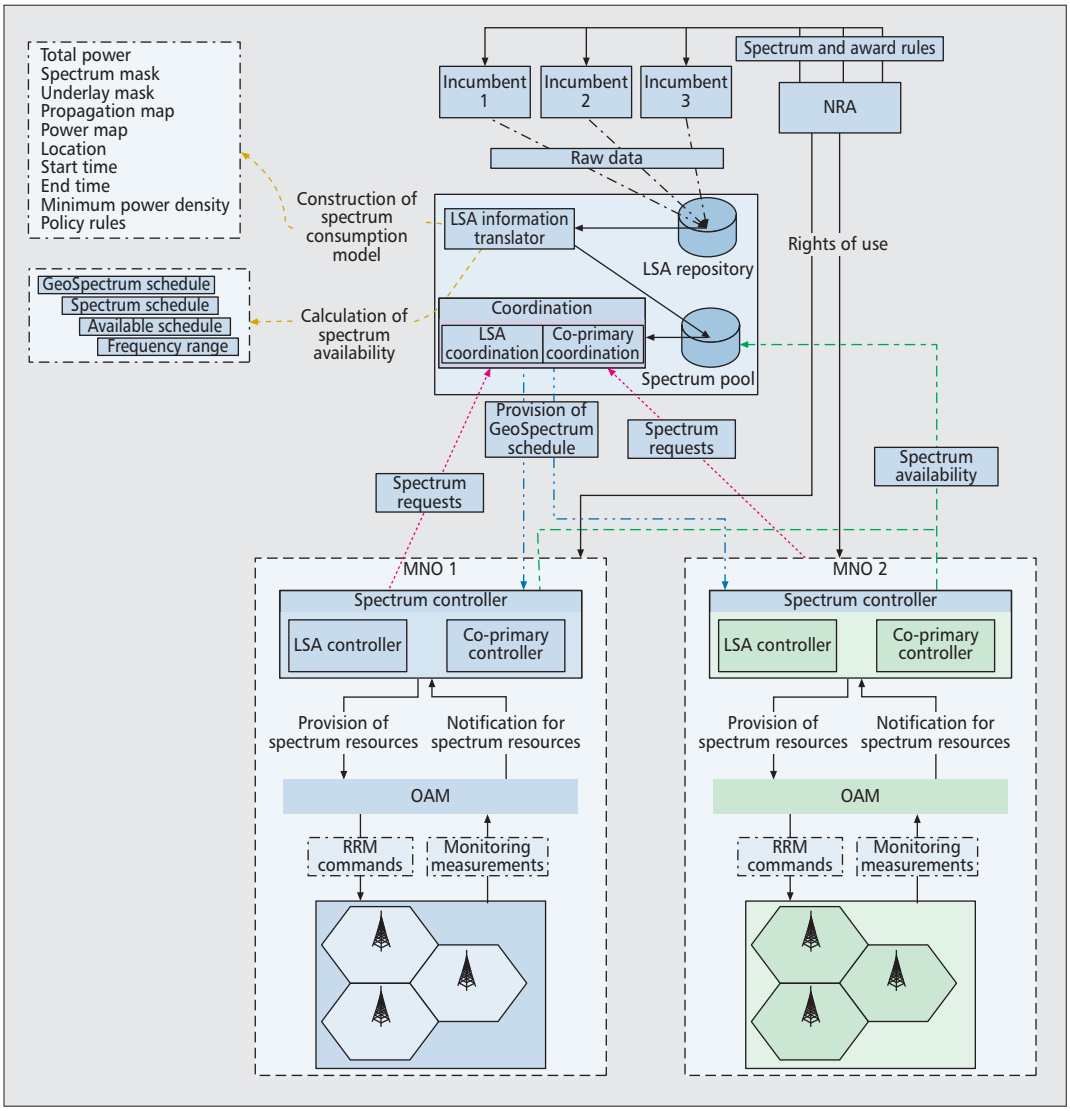


Figure 1. Combined co-primary and LSA shared access functional architecture.

The MNOs periodically inform the spectrum pool regarding their available spectrum resources so as to enable the spectrum sharing process. The information to be provided is related to spectrum availability with regards to time, space, and frequency domains.

Radio resource management (RRM): It provides monitored information (e.g., congestion level, blocking rate) to the OAM and receives information from it about any newly allocated spectrum chunk.

The following section describes the information being exchanged among the above-described entities for enabling the spectrum sharing framework.

Information Exchange Structures: The MNOs periodically inform the spectrum pool regarding their available spectrum resources in order to enable the spectrum sharing process. The information to be provided is related to spectrum availability with regard to time, space, and frequency domains. Such information is formed as a spectrum availability information data structure similar to the structures of Protocol to Access White Spaces (PAWS) for building geolocation databases to exploit TVWS [15]. Figure 2 depicts the composition of this data structure. More specifically:

- Spectrum availability information is a list of geospectrum Schedules.
- A geospectrum schedule incorporates location information and a list of spectrum schedules for each location, thus implying availability over space for the associated list of spectrum schedules.
- A spectrum schedule indicates availability over time (i.e., parameters start time and end time) for a set of spectrum chunks (highlighted as available spectrum data structure).
- Spectrum chunks are characterized by the operating bandwidth and a list of frequency ranges to indicate the spectrum chunks' availability over frequency bands.
- Frequency range is characterized by boundaries (i.e., lower frequency and higher frequency parameters) and the maximum effective isotropic radiated power (EIRP) density.

The spectrum availability information is calculated in the coordination entity based on the data aggregated to the spectrum pool from license holders (either MNOs or other incumbent users). This information is being provided from the coordination entity to the spectrum controller entity of the MNO and from there to the OAM.

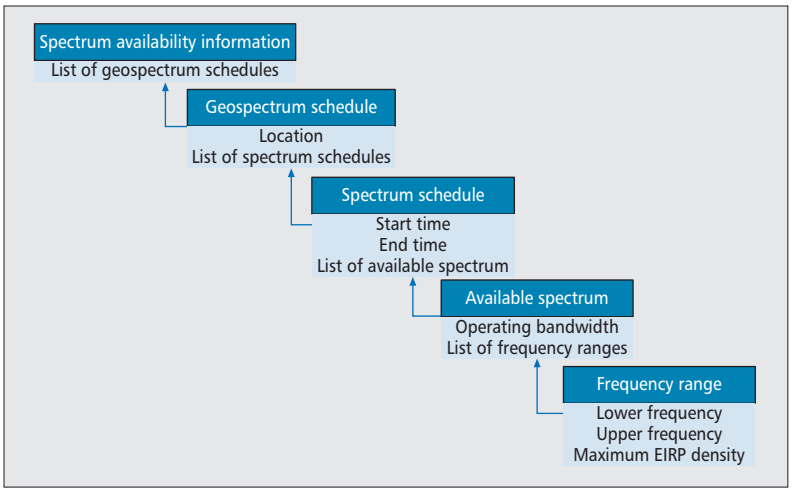


Figure 2. Spectrum availability information structure.

Spectrum consumption model data constructs	
Data construct	**Description**
Total power	Defines the total power on the radio side of the antenna and it serves as the reference value for relative power values found in other constructs.
Spectrum mask	Specifies spectral power density of a transmitted signal, i.e. the spectral bounds of the signal
Underlay mask	Specifies the spectral power density relative to a maximum power density at a receiving antenna of the maximum allowed interference by a remote interfering transmission.
Propagation map	Specifies a rate of attenuation by direction. This rate is specified using the exponent of the log-distance pathloss model.
Power map	Specifies the directional values of the transmitted power density relative to the total power of the model.
Location	Conveys the location or region where RF components (transmitters/receivers) may be.
Start time	The start time is the time that a model begins to apply. The start time may also be used to define a periodic use of spectrum that begins with the start time.
End time	The end time is the time that a model stops applying.
Minimum power density	The minimum power density specifies the attenuation level where transmitted signals are no longer protected. It is used as part of a transmitter model, usually when modeling a broadcasting service where reception is based only on the range of the transmission.
Policy rules	Rules that define the behavior of LSA licensee (e.g., how often he queries the repository, rules based on rights of use, interference)

Table 1. Spectrum consumption model data structures applicable to the LSA spectrum sharing case.

A fundamental requirement for the proposed system is not to burden the initial license holders with complex calculations. In the LSA case the incumbent users provide reports relevant to their configuration and the way they use their acquired spectrum. Such information cannot be used directly by the MNOs, who need to know the time/frequency/location domain in which spectrum resources are available. Thus, a translation process is required to transform the spectrum consumption characteristics to a spectrum availability schedule such as the one described above. The LSA information translator undertakes this task and incorporates the spectrum availability information to the spectrum pool. The information required from the LSA information translator for building the spectrum availability information is aggregated to the LSA repository based on inputs from the incumbent users. The type of information that our proposed system shall collect could be summarized in the following list:

- Spectrum occupancy characteristics: the spectrum mask together with space (i.e., location) and time (i.e., start time and end time) details that defines the spectrum occupied by an incumbent user
- Power characteristics: the total power and power map that determine the incumbent user's transmission characteristics
- Propagation characteristics: suggests the propagation map and minimum power density that determine the attenuation level
- Geographical area information: conveys the region where RF components exist
- Time details: consists of the start time and end time that determine the time period during which the incumbent user is using the spectrum resources with the aforementioned characteristics

The previous information may be further decomposed to a set of data constructs. Table 1 summarizes the data constructs required from the LSA information translator for the extraction of spectrum consumption model of each incumbent user.

Fuzzy Logic Based Spectrum Allocation

The previously described architecture enables the MNOs to discover information about any spectrum opportunities and proceed in agreements for sharing spectrum. However, the way that each MNO decides which spectrum is more suitable for covering its needs has yet to be defined. Several proposals exist in the literature based on game theoretic approaches [6], cooperative interference compensation [7], and so on, but do not take into account the MNOs' special needs regarding network load trend, user requirements, user mobility, and so on. The diversity of these inputs and the complexity of the problem make it hard to handle in real-time timeframes. Thus, we propose the introduction of a reasoning scheme, based on fuzzy logic controllers, to identify the most suitable spectrum for covering an MNO's needs in a specific location and time. Fuzzy logic is an ideal tool when dealing with multi-variable problems with fuzzy inputs and often contradic-

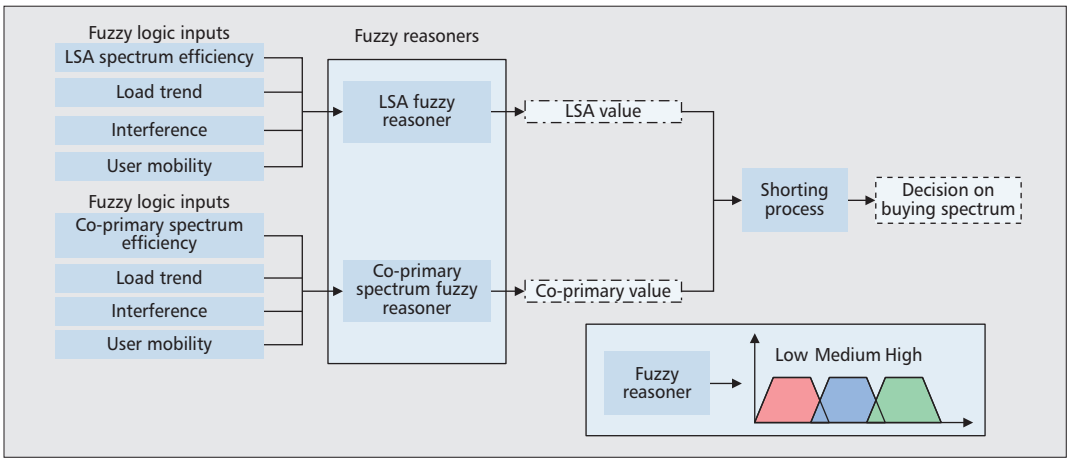


Figure 3. Fuzzy Logic based spectrum controller.

The LSA sharing scheme is not interference free compared to the co-primary sharing scheme. Furthermore, spectrum provided under LSA sharing scheme is time/location/frequency specific, while in case of co-primary such limitations lie on the mutual agreements among the MNOs.

tory objectives. The fuzzy logic controller incorporates the operator's renting strategy to maximize its revenues while covering the users' needs.

Fuzzy Logic Enhanced Spectrum Sharing Algorithm

A fuzzy logic controller (FLC) consists of three parts: the fuzzifier, the inference system, and the defuzzifier. The first part is responsible for mapping (fuzzifying) the input values to the extent that these values belong to a specific state (e.g., low, medium, high using the input membership functions). The second part (inference system) uses simple "IF … THEN…" rules to identify the relation of the inputs to the outputs; each rule applies to a certain degree for every output. Then the output degrees for all the rules of the inference phase are aggregated (using the output membership functions). The actual output of the decision making process comes from the defuzzification procedure, which captures the decision of the decision maker (i.e., proceed in this action, e.g., buy LSA spectrum chunks).

In the proposed spectrum sharing scheme we introduce an algorithm consisting of two fuzzy reasoners, one for co-primary and one for LSA spectrum sharing authorization options. The algorithm resides in the spectrum controller block of each MNO (Fig. 1). Each fuzzy reasoner captures the corresponding suitability of each sharing scheme to fulfill future demands. Each fuzzy reasoner takes into consideration four inputs (Fig. 3). Three of them are related to network conditions and are the same for both fuzzy reasoners: load trend, interference levels, and average user mobility. The fourth one captures the spectrum efficiency of each sharing scheme. More specifically, the considered fuzzy reasoner inputs are the following:

- Load trend: captures the (overall) user bandwidth demands trend over a time window. High load trend indicates that the network capacity may be insufficient to serve user demands in the near future, thus showing the increasing need to buy spectrum to fulfill user needs.
- Interference level: captured by the average packet error rate in a specific geographical area.
- Average user mobility: captures the dynamic nature of the way users affect the traffic load in a specific location due to their mobility behavior and is defined static (users moving from 0 to 3 km/h), semi-static (users moving from 3 to 10km/h), and high moving users (for speeds above 10 km/hr).
- (LSA or co-primary) spectrum efficiency: relates the past transactions to future spectrum requests. Specifically, if an operator has not exploited the previously acquired spectrum, future transactions for buying new spectrum become unattractive.

The inference engine of each of these fuzzy reasoners produces a suitability factor indicating whether LSA or co-primary spectrum resources should be obtained. Each fuzzy logic controller evaluates the inputs in a different manner, given the fact that each spectrum authorization option has its own key characteristics. The LSA sharing scheme is not interference-free compared to the co-primary sharing scheme. Furthermore, spectrum provided under the LSA sharing scheme is time-/location-/frequency-specific, while in the case of co-primary such limitations lie ing the mutual agreements among MNOs. Each fuzzy logic controller incorporates 54 rules for combining the inputs to a suitability output. The rules capture the strategy of the operator, which is developed based on the characteristics of each sharing scheme. LSA is time-/geographic-area-specific, whereas co-primary spectrum sharing is time-/geographic-area agnostic, thus making it more suitable for highly moving users. Furthermore, given the fact that the LSA licensees are not necessarily coordinated, interference may occur among them, which will increase the packet error rate. Thus, it is less preferable than interference-free co-primary spectrum sharing in high interference areas. Finally, it is assumed that the operator wishes to consume the already acquired spectrum before proceeding in spectrum requests.

Based on the nature of each input, we have used several types of input membership functions for capturing the special characteristics of each input. More specifically, for the average user mobility and the load trend, we have used triangular membership functions because at certain values we are certain about the state they

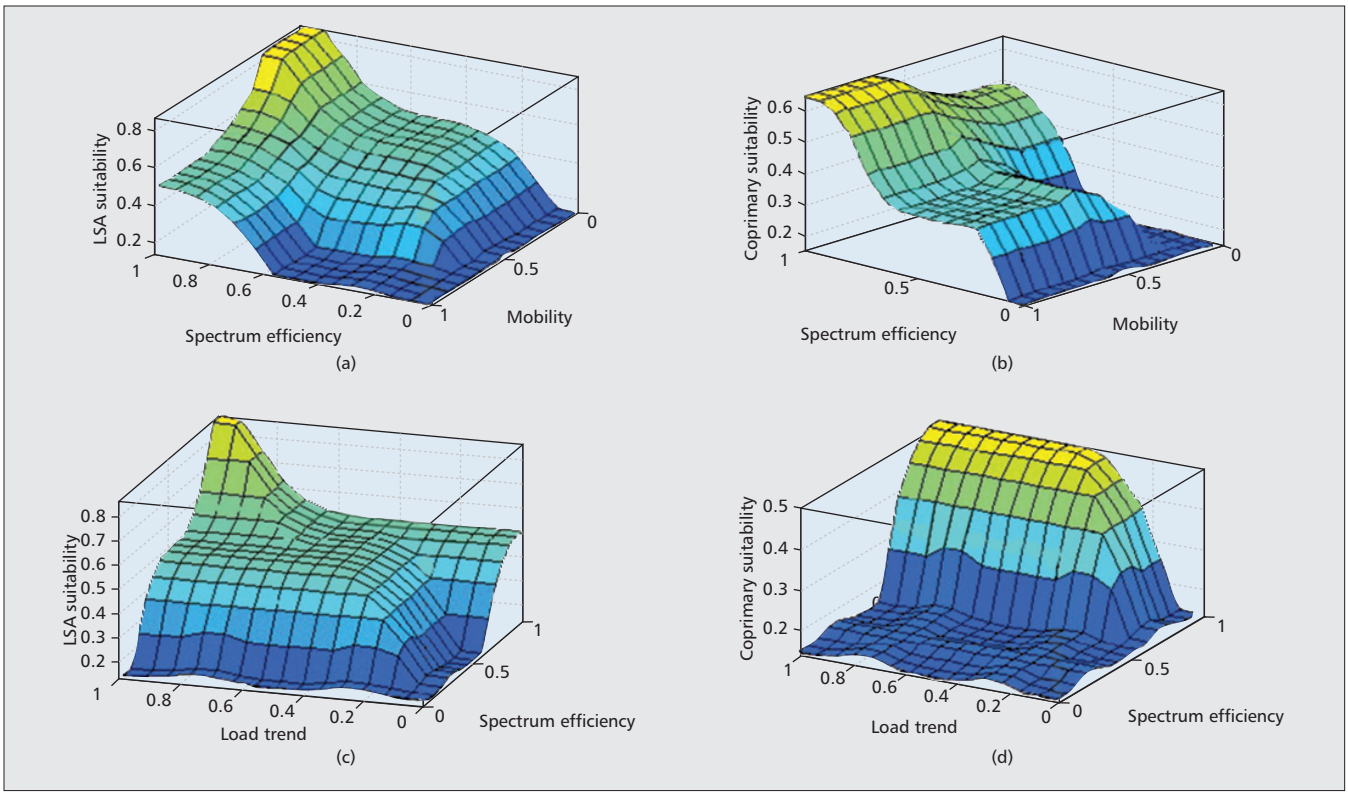


Figure 4. LSA and co-primary spectrum sharing suitability: a) LSA suitability in relation to spectrum efficiency and average user mobility; b) co-primary suitability in relation to spectrum efficiency and average user mobility; c) LSA suitability in relation to load trend and spectrum efficiency; d) co-primary suitability in relation to load trend and spectrum efficiency

are capturing (e.g., low/high mobile user, low/high load trend). On the other hand, for the interference and spectrum efficiency we have used Gaussian membership functions to exploit the non-zero nature of this membership function at the definition domain. Regarding the output, we are also using Gaussian membership functions for their smoothness in the decision making process. Since the LSA sharing scheme imposes constraints over the time/frequency/location domains and enables spectrum usage from multiple licensees, it is assumed in our model that acquiring LSA spectrum resources will be less expensive than co-primary sharing.

Fuzzy Logic Enhanced Spectrum Sharing Analysis

As described in the previous section, the fuzzy reasoners of co-primary and LSA spectrum sharing strategies are configured differently so as to capture the special characteristics of each approach. Thus, depending on the input values, the outputs will differ and the decision maker will conclude to the one approach that is most suitable to the network context. Figure 4 presents the suitability values for both fuzzy reasoners in relation to the inputs which is the outcome of our experimentation for the validation of the fuzzy logic controllers. The top part of the figure (Fig. 4a and 4b) presents the ranges of the suitability in relation to the spectrum efficiency and average user mobility having as parameters the interference and the load trend (interference is set to 0.5 indicating acceptable QoS conditions for the served users and load trend is set to 0.9, indicating very increasing load demands). We may observe that (Fig. 4a) LSA becomes attractive if the average user mobility is low and the acquired spectrum tends to be insufficient for covering the user requirements. On the other hand, co-primary spectrum sharing (Fig. 4b) is more attractive under the same conditions when the average user mobility is high.

Similarly, the bottom part of the figure (Figs. 4c and 4d) presents the ranges of suitability in relation to the load trend and spectrum efficiency, having as parameters the interference level and average user mobility (interference is set to 0.5, indicating acceptable QoS conditions, and the average user mobility is set to 0.1, indicating low average mobility). Figures 4c and 4d show that LSA and co-primary become attractive (i.e., suitability factor is increased) when the load trend increases. However, LSA is more attractive (i.e., suitability factor ranges from 0.5 to 0.8 in most cases) compared to the co-primary sharing option (i.e., maximum suitability factor is 0.5). This is reasonable due to the low mobility assumed and the assumption, based on which our fuzzy logic rules were built, indicating that LSA is a less expensive option compared to co-primary sharing.

Conclusion

Innovative approaches are required to cover the augmented requirements of future networks to reduce spectrum resources shortage. Considering that other frequency bands remain underuti-

lized due to limited data transmissions of their rightful users, the exploitation of such bands becomes very attractive. Up to now, exploitation of TVWS in cognitive radio approaches has been considered, although the drawbacks (i.e., complex solutions, interference may be caused to the mobile user, etc.) of such solutions still discourage their introduction. On the other hand, the rise of new approaches, such as co-primary spectrum sharing and licensed shared access, which protect both spectrum license holders and spectrum licensees, enable flexible spectrum management.

In this article the vision of future mobile networks in which the MNOs share spectrum resources with either other MNOs (co-primary sharing scheme) or incumbent users (LSA sharing scheme) has been thoroughly presented by describing the key characteristics of each approach. The analysis could be summarized in the two main differences between these two spectrum sharing approaches. Both of the differences are related to actors involved in the sharing procedure. The first one is related to the incumbent users that shall not be burdened with complex calculations, which implies that in the LSA case the presence of a translation and coordination entity is required. The second main difference is related to the fact that in the LSA concept several spectrum licensees (i.e., MNOs) may exist, which will not necessarily be coordinated; this may introduce interference among them (the incumbent user is protected from interference), whereas in co-primary spectrum sharing the spectrum buyer will not experience interference for the time period of the rental. In our work we have presented a common architectural framework for coupling the co-primary and LSA sharing schemes. For meeting the requirement of reduced complexity in the incumbent users, we propose the introduction of a translation engine, with the prerequisite that the data will be formed in spectrum availability structure indicating available spectrum over time, frequency, and geographical domains.

The proposed architecture is accompanied by a fuzzy-logic-based spectrum sharing algorithm for enabling - operators to decide which spectrum authorization option is more suitable given the network conditions. In our analysis, fuzzy reasoners only for the LSA and co-primary sharing schemes have been presented. However, the proposed algorithm could easily be extended to other sharing schemes (e.g., general authorization schemes). Finally, as the expenses of buying (or renting) additional spectrum are related to the market demands, the spectrum sharing scheme is planned to be further extended with adaptation mechanisms (i.e., reinforcement learning techniques) to tune the decision making process.

Acknowledgment

This work has been performed in the framework of the FP7 project ICT-317669 METIS. The authors would like to acknowledge the contributions of their colleagues. This information reflects the Consortium's view, but the Consortium is not liable for any use that may be made of any of the information contained herein.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018," Feb. 5, 2014, http://www.cisco.com/assets/sol/sp/vni/forecast_highlights_mobile/index.html.
[2] ITU M.2072-0, 2006, "World Mobile Telecommunication Market Forecast," Milan, Italy, 2006.
[3] ITU-R. R. M.2243, "Assessment of the Global Mobile Broadband Deployments and Forecasts for International Mobile Telecommunications," 2011.
[4] J. Wannstrom, "LTE-Advanced (2012)," May 10, 2012, http://www.3gpp.org/IMG/pdf/lte_advanced_v2.pdf.
[5] K. Chatzikokolakis *et al.*, "Spectrum Aggregation in Cognitive Radio Access Networks: Business and Power Control Aspects," submitted for publication in "Evolution of Cognitive Networks and Self-Adaptive Communication Systems," *IGI Global*, 2012.
[6] Y. T. Lin, H. Tembine, and K. C. Chen, "Inter-Operator Spectrum Sharing in Future Cellular Systems," *2012 IEEE GLOBECOM*, 2012, Dec., pp. 2597–2602.
[7] G. Salami and R. Tafazolli, "Interoperator Dynamic Spectrum Sharing (Analysis, Costs and Implications)," *Int'l. J. Computer Networks (IJCN)*, vol. 2, 2010, pp. 47–61.
[8] A. Apostolidis *et al.*, "Intermediate Description of the Spectrum Needs and Usage Principles," Project METIS deliv. D5.1, https://www.metis2020.com/.
[9] M. Bennis and J. Lilleberg, "Inter-Network Resource Sharing and Improving the Efficiency of Beyond 3G Systems," *41st Annual Conf. Info. Sci. and Sys.*, 2007, 14–16 Mar. 2007, pp. 357, 362.
[10] H. Kamal, M. Coupechoux, and P. Godlewski, "Inter-Operator Spectrum Sharing for Cellular Networks Using Game Theory," *IEEE PIMRC '09*, 13–16 Sept. 2009, pp. 425–29.
[11] SAPHYRE D4.1 "Resource Allocation and Interference Management Strategies," Dec. 2011.
[12] RSPG Opinion on Licensed Shared Access, Nov. 2013, ref. RSPG13-538.
[13] ECC report 205 Licensed Shared Access (LSA), CEPT, Feb. 2014.
[14] ETSI TS 103 154, "Reconfigurable Radio Systems (RRS); System Reference Document; Mobile broadband Services in the 2300 MHz–2400 MHz Frequency Band Under Licensed Shared Access Regime."
[15] "Protocol to Access White-Space (PAWS) Databases: Use Cases and Requirements," IETF Internet draft, Dec. 2013.

Biographies

Konstantinos Chatzikokolakis (kchatzi@di.uoa.gr) received his B.Sc. and M.Sc. from the Department of Informatics and Telecommunications of the National Kapodistrian University of Athens in 2008 and 2012 respectively. He is currently a Ph.D. candidate at the same department and a member of SCAN Lab since 2012. He has participated in several European projects including SACRA, Univerself, and METIS. His research interests include dynamic resource allocation, cognitive radio networks, network management, spectrum management, and spectrum sharing.

Panagiotis Spapis received his diploma in electrical engineering in 2008 and his Ph.D. in telecommunications in 2015. He worked as a researcher in the Department of Informatics and Telecommunications of the University of Athens, participating in several EU projects and publishing parts of his work in numerous journals and conferences. Since June 2014 he is a researcher in Huawei's ERC (Munich) 5G RAN group. His research interests lie in the areas of situation and context awareness in 5G systems.

Alexandros Kaloxylos has participated in numerous EU projects and has published over 100 papers in the area of mobile communications since 1994. He is an Editorial Board Member of *IEEE Surveys and Tutorials*, and a TPC member for numerous conferences. He is an assistant professor in the Department of Informatics and Telecommunications of the University of Peloponnese. Since June 2014, he has been a principal researcher in Huawei's ERC (Munich), where he is currently a team leader for 5G RAN.

Nancy Alonistioti has a B.Sc. degree and a Ph.D. degree in informatics and telecommunications. She has participated in several national and European projects (e.g., E3, UNIVERSELF, SACRA, METIS). She served as a lecturer at the University of Piraeus and recently joined the faculty of the Department of Informatics and Telecommunications of the University of Athens. She has over 100 publications in the area of mobile communications, reconfigurable, cognitive, and autonomic systems and networks, and future Internet.

As the expenses of buying (or renting) additional spectrum are related to the market demands, the spectrum sharing scheme is planned to be further extended with adaptation mechanisms (i.e. reinforcement learning techniques) to tune the decision making process.

Coordination Protocol for Inter-Operator Spectrum Sharing in Co-Primary 5G Small Cell Networks

Bikramjit Singh, Sofonias Hailu, Konstantinos Koufos, Alexis A. Dowhuszko, Olav Tirkkonen, Riku Jäntti, and Randall Berry

ABSTRACT

We consider spectrum sharing between a limited set of operators having similar rights for accessing spectrum. A coordination protocol acting on the level of the RAN is designed. The protocol is non-cooperative, but assumes an agreement to a set of negotiation rules. The signaling overhead is low, and knowledge of a competitor's channel state information is not assumed. No monetary transactions are involved; instead, spectrum sharing is based on a RAN-internal virtual currency. The protocol is applicable both in a scenario of mutual renting and when the operators form a spectrum pool. The protocol reacts to variations in interference and load of operators, and shows gains in a simulated small cell scenario.

INTRODUCTION

In state-of-the-art mobile communication, dedicated and exclusive spectrum access coupled with unlicensed local area solutions is the mainstream approach that national regulatory authorities use to allocate new spectrum. In exclusive access, only one operator has the right to use a dedicated licensed frequency band according to specific rules. Although exclusive spectrum access will certainly be needed in fifth generation (5G) mobile systems to guarantee quality of service (QoS) in wide area radio access networks (RANs), other regulatory options may be needed in addition. Especially at carrier frequencies above 6 GHz and in small cell networks, exclusive access may result in low spectrum utilization efficiency. Unlicensed access, on the contrary, offers unpredictable QoS. Enabling the high capacity and flexible usage envisioned for 5G systems thus calls for more flexible regulatory regimes where, for example, operators may use various frequency bands with different authorization modes [1].

Co-primary shared access is a complementary new alternative for spectrum sharing, where multiple operators jointly use a part (or all) of their licensed spectrum [1, 2]. The most relevant co-primary shared access scenarios are mutual renting (MR) and limited spectrum pool (LSP). In MR, operators have individual licenses to access exclusive frequency bands, and are mutually allowed to *rent* parts of their licensed resources to their peers upon request. In LSP, a group license is given to an operator for using a common pool of spectral resources, which is shared with a limited set of operators that have equal access rights. As the set of peer operators and the principles of spectrum usage are known beforehand, investment decisions under co-primary shared access have lower risk because the long-term share of resources can have a predictable minimum value [2].

Joint use of licensed spectrum among operators can be realized either orthogonally in time [3], frequency [4, 5], or space, or non-orthogonally [5, 6]. In cooperative orthogonal time domain sharing, operators with a low load can lend their time slots to heavily loaded operators, helping them to reduce blocking probability and frame delay [3]. An upper bound for the sum capacity of a two-operator orthogonal frequency domain sharing scenario is found in [4], where operators have full access to a user-equipment (UE)-specific channel quality indicators of all shared channels and perform coordinated scheduling. Orthogonal sharing based on pairwise exchange of resource blocks between two operators is considered in [5].

In non-orthogonal spectrum sharing, operators simultaneously use a common block of spectral resources, creating inter-operator interference. In [6], a cooperative game approach was proposed that converges quickly to a value close to the Nash bargaining solution; however, it requires full knowledge of action profiles in the neighborhood of the node. Non-orthogonal inter-operator spectrum sharing in the spatial domain was considered in [5], where

Bikramjit Singh, Sofonias Hailu, Konstantinos Koufos, Alexis A. Dowhuszko, Olav Tirkkonen, and Riku Jäntti are with Aalto University.

Randall Berry is with Northwestern University.

0163-6804/15/$25.00 © 2015 IEEE

channel state information (CSI) is exchanged among operators to implement coordinated transmit beamforming by steering their antenna beams in the most convenient direction. In all these studies, the operators benefit from cooperative spectrum sharing; however, they need to reveal proprietary information to their competitors [5, 6] or to a central entity [3, 4].

Operators are competitors by nature; therefore, the rationale to make them cooperate would be either a legal framework or self-interest. In such settings, there is no reason to assume that operators are willing to exchange proprietary information with their competitors. Moreover, operators provide differentiated services to their customers, with objectives that can be categorically different according to their business models. Thus, the optimization of a joint utility by a central entity is not realistic. Co-primary spectrum sharing between operators is thus characterized by the operators not knowing each other's optimization target or network states.

Spectrum sharing can be realized by monetizing spectrum usage and arranging auctions to determine spectrum management [7]. Here we are interested in adaptive spectrum use on a short timescale (e.g., related to cell load variations and changing inter-operator interference conditions). It is a nontrivial task to design efficient auction mechanisms for a limited area for a limited time, and to couple operator auction strategies to their income model. Implementing auctions also requires the involvement of a trusted third party with substantial accounting infrastructure to collect bids, track payments, and so on. This adds inertia in moving toward auction-based spectrum sharing. Accordingly, we consider non-monetized spectrum sharing directly between the RANs of the operators.

The interaction between self-interested players can be modeled by non-cooperative games, where players make decisions independently. A non-cooperative one-shot game formulation for unlicensed access was discussed in [8]. The players were not constrained by any rules, and freely selected power allocation strategies. When the inter-link interference is sufficiently low, the flat power allocation over all available channels represents the unique Nash equilibrium (NE) for one-shot spectrum sharing games with complete opponent information. Multi-operator spectrum sharing differs from this setting, as the utilities and strategies are player-specific and not shared among players. There may also be a set of rules, either agreed among players or enforced by a legal entity. The strategic choices of the players are bound by these rules; no deviations from the rules are allowed. For instance, the legal framework governing the operation of different systems in license-exempt bands (e.g., WiFi and Zigbee) gives equal access rights to all radio devices when complying with certain power emission levels, spectral masks, channel reservation protocols, and activity rates in the case of industrial, scientific, and medical (ISM) bands. Considering long-term sharing in license-exempt bands, repeated game strategies that lead to favorable NE were also discussed in [8]. These games have no a priori rules. It is assumed that nodes exchange information and agree in advance on their operational points in terms of a power allocation across the shared frequency channels. The agreement is then enforced under the threat of punishment. Obviously, without exchanging information about network states and optimization objectives, it is not possible to identify and punish cheating, and cheating as such becomes ill defined. Thereby, in a co-primary spectrum sharing setting, the applicability of strategies based on concepts of dishonesty and punishment may be limited.

Unlike spectrum sharing in license-exempt bands, where the number of players sharing spectral resources is indefinite, we consider a fixed and known number of players with publicly known and persistent identity. We assume no intra-RAN information exchange, except for limited message exchange among operators to realize spectrum negotiations. This requires a new interface, which may be over the air or over the core network. We design a coordination protocol for inter-operator spectrum sharing that incorporates both operator strategies and spectrum sharing rules. The proposed protocol does not require a priori agreement about the operational points of different operators, but it leads to operational points that are better for all operators compared to the case without spectrum usage coordination.

As a first assumption we assume that an operator ignores the state of the other operator when negotiating the use of spectrum. The decision on spectrum usage of an operator is based on its own network state, such as network load, channel conditions, UEs' locations, and interference caused by the opponent operator.

System Model

This article considers co-primary spectrum sharing among a limited number of co-located RANs belonging to different operators. The operators' RANs are full cellular networks with multiple base stations (BSs). A coordination protocol tailored for spectrum sharing in small cell networks is considered. At least in the first stage, it is not likely that macrocells will participate in spectrum sharing. Only a part or cluster of the RAN may be involved in spectrum sharing (e.g., small cells of different operators located in the same or neighboring buildings). Small cells of different operators may offer high data rate services in different buildings, or have partially or fully overlapping service areas.

For simplicity, we concentrate on a two-operator scenario. The operators are self-interested and not willing to share operator-specific information such as load, channel usage, or CSI. An operator is unable to reliably estimate its opponent's optimization targets and network load from RAN measurements. As a first assumption we assume that an operator ignores the state of the other operator when negotiating the use of spectrum. The decision on spectrum usage of an operator is based on its own network state, such as network load, channel conditions, UEs' locations, and interference caused by the opponent operator. Operators divide their spectrum into equal-size component carriers (CCs), and contribute to the spectrum sharing algorithm with an equal number of CCs. The transmit power that is used per CC is assumed constant, and spectrum sharing in the downlink is considered. This enables reliable estimation of interference caused by another operator.

We adopt a simplified version of the European Telecommunications Standards Institute

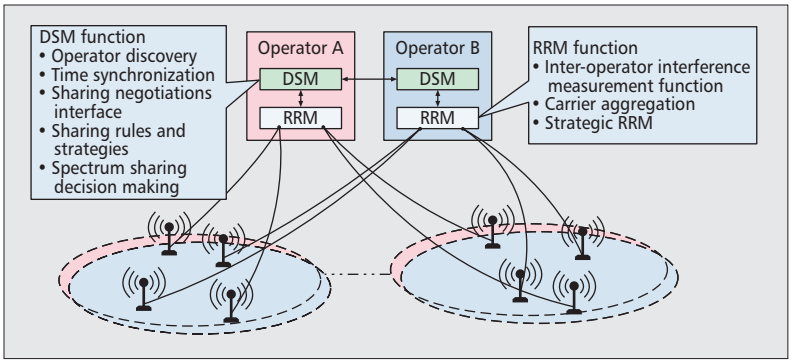


Figure 1. Elements in system functional architecture required for inter-operator spectrum sharing on the RAN level.

(ETSI) reconfigurable radio systems (RRS) functional architecture [9] for the considered co-primary spectrum sharing (Fig. 1). The dynamic spectrum management (DSM) block is responsible for the short- and long-term management of spectrum. It decides spectrum usage and is engaged in the discovery of other operators by contacting external spectrum databases or repositories. The operators communicate only through their DSM. Coarse time synchronization is needed so that operators indicate their spectrum sharing proposals and decisions to each other almost simultaneously.

The radio resource management (RRM) block performs intra-RAN interference mitigation to utilize the limited RF resources as efficiently as possible. In addition, it supports inter-operator spectrum sharing by providing inter-operator interference measurements, carrier aggregation, and strategic RRM for negotiating spectrum between operators.

Coordination Protocols

A coordination protocol is a mechanism to handle spectrum sharing negotiations between peer networks in inter-operator spectrum sharing. Such a coordination protocol requires:

- A peer-to-peer connection between the operators so that each operator can indicate its spectrum sharing proposals to the others
- A set of decision rules determining the sharing outcome based on the operator proposals. These rules may, for instance, be agreed in advance between operators or enforced by an external regulator.

Spectrum Sharing Based on Spectrum Usage Favors

We assume that operators make proposals about how to share spectrum considering only their own interest, while also respecting a set of spectrum sharing rules. A priori, the operators would be required to follow a specific MAC protocol. For simplicity, we assume that operators involved in LSP have by regulation equal access rights over the pool, and an operator should always have the right to use the full spectrum pool if desired. We shall see that with a suitable coordination protocol, a rational operator may not always use the full spectrum pool. In MR, on the other hand, each operator has the legal right to access its own spectrum, and may also give rights to other operators to use it. Thus, under MR, an operator is entitled to use its own spectrum exclusively if it so desires. Hereafter, the state that an operator can take at any time instant without breaking any rules is referred to as the *fallback state*.

The coordination protocol is essentially a mechanism enabling the operator to determine the self-optimal way to use the spectrum at a particular time and announce it to the opponent. The opponent has the legal right to accept or reject the received proposal for spectrum usage. Any proposal that is accepted by the opponent necessitates a departure from the fallback state.

With concave utility functions, an accepted spectrum sharing proposal provides instantaneous gain in the network utility of the operator that made the proposal and reduces the instantaneous network utility of the opponent operator. An accepted proposal resulting in a deviation from the fallback state of the underlying MAC protocol can be seen as a *spectrum usage favor* because the opponent is aware of its utility loss. A spectrum usage favor is exchanged only if one operator asks for it and, simultaneously, the other operator is willing to grant it. The operators are not forced to act. Spectrum sharing based on spectrum favors entails the benefits of opponent-blind operation, non-monetized spectrum use, and a few bits of inter-operator signaling per protocol time unit.

Obviously, operators are not willing to accept everlasting performance loss in their network utilities from granting favors. The spectrum favors should be valid for a certain time period, agreed a priori between operators, and should be a part of the spectrum sharing rules in the coordination protocol. After the specified time period expires, the resource allocation falls back to the state in which it was before granting the favor.

While the operator taking a favor gets instantaneous improvement in its network utility, it is not immediately clear what the benefit is for the operator that grants a favor. As operators are self-interested entities, they will not give favors for free. Since we do not consider inter-operator monetary transactions, the benefits from granting a favor should lie in reciprocity. It is well known from a game theoretic perspective (e.g., in the tit-for-tat strategy for the repeated prisoner's dilemma) that a player will be cooperative if and only if the opponent cooperates in return. A form of reciprocity could be, for instance, that both operators give and take an equal amount of equally valuable spectrum favors.

We discuss two coordination protocols that can be distinguished based on the reciprocity time horizon. First, we consider the case of impatient operators that care only about instantaneous benefits. Then we focus on patient operators that are interested in long-term benefits. In both cases, we assume non-cooperative stage games played in sequence, with fixed action spaces and rules. The games are stochastic because the rewards at each stage depend on parameters governed by probability distributions,

such as the network load, UE deployment, and channel fading states. Rewards are computed with respect to the fallback state of the underlying spectrum sharing scenario.

Instantaneous Reciprocity

When the operators are impatient, we can view each stage as a separate one-shot game in which reciprocity must be satisfied. Thus, each operator gives and takes an equal number of favors at each stage game. Since the operators are selfish, a favor is exchanged only if both operators experience a positive reward at that stage.

Given the underlying spectrum sharing scenario, a set of strategies is available for the operators. The strategy is essentially the type of spectrum usage favor asked for by the operators. The target is to design strategies and spectrum sharing rules that result in favorable NE solutions.

In MR, the strategy may correspond to the amount of its own licensed spectrum the operator would like to share with the opponent, whereas in LSP, the strategy may correspond to the fraction of the pool the operator is willing not to use. Given the spectrum sharing scenario, the operators implement their strategies independently and exchange their proposals with the other players. After that, we need a suitable rule to resolve the spectrum sharing proposals submitted by the operators. One possibility is to select the minimum of them. Following this minimum rule, no operator will share more spectrum than desired under MR, whereas no operator will vacate more spectrum than desired under LSP. As a result, when an operator does not benefit from departing the fallback state, no favor is exchanged.

For concave utilities, the one-shot game following the proposed minimum rule and strategies is characterized by a unique NE [10]. In [8], spectrum sharing in the power domain was considered, and it was shown that the NE in a one-shot game is trivial with equal power on all carriers. Here, the game takes place in frequency resources, and the resulting NE may depart from full sharing. The outcome approaches the result of the cooperative games of [6].

Long-Term Reciprocity

Operators are expected to share spectrum for a long time. As an operator has a persistent and publicly known identity, the operators can learn from each other's behavior. Accordingly, the interaction between operators can be modeled as a repeated non-cooperative game. In the repeated game, we need to keep a book of the favors exchanged because past rewards do impact future decisions. For symmetric operators, it is natural to assume that under long-term reciprocity, operators should give and take the same amount of equally valuable favors over some specified time horizon (or equivalently with some discount factor). The time horizon depends on the level of patience that operators have. Here, we consider infinitely patient operators.

In a sequence of repeated interactions over a long-time horizon, self-interested operators with no information about the RANs of other operators can develop methods to ensure gains from coordination. For instance, operators may take advantage of uncorrelated load variations in their RANs. Then an operator with a high instantaneous load may get spectrum usage favors from an opponent operator if the opponent happens to have low load. In the future, this operator will have the chance to return these favors to maintain reciprocity. Even though an operator experiences performance loss by granting a favor at a stage of the repeated game, when the load situation changes, it may get a performance gain that outweighs its past loss. While exchanging favors, each operator ensures that its expected gain is larger than its expected loss, thus benefiting when compared to the case where no favors are exchanged.

In different spectrum sharing scenarios, different kinds of spectrum usage favors are asked and granted by operators. In LSP, we view a single type of favor: an operator asks the opponent for permission to exclusively use some resources from the pool. In MR, there can be bilateral agreements for resource utilization. In that case, there are two types of favors:

- An operator asks the opponent for permission to start jointly using some of the resources of the opponent.
- An operator asks the opponent for permission to start exclusively using some of the resources of the opponent.

Repeated games admit a large set of equilibrium points. Since the considered game is stochastic due to the time variation of network states, it is hard to analyze and find its NE. To proceed with analysis, we resort to heuristic strategies attempting to obtain long-term reciprocity. This can be done with a threshold-based approach.

We assume that an operator knows the probability distribution functions of its utility gains and losses. A simple long-term reciprocal strategy would be that an operator asks for a favor if its immediate utility gain is higher than a threshold θ_g, and grants a favor (upon being asked) if its immediate utility loss is smaller than another threshold θ_l. These decision thresholds depend not only on the current network state and the gain and loss statistics, but also on the sequence of previous interactions with opponent operators. The thresholds for operator A would be coupled to the probabilities of operator B to grant and ask favors so that long-term reciprocity is achieved. Details for setting the decision threshold in the case of an LSP can be found in [11]. The proposed heuristic strategy performs strictly better than the strategy that does not involve the exchange of favors. Unlike the one-shot game, where an operator takes and gives favors simultaneously, we simplify the repeated game by assuming that no action is taken when both operators ask a favor at the same stage of the game.

Combined Coordination Protocol

In co-primary small cell deployments, the interference conditions and the RAN load are expected to vary significantly. When UEs are located close to their serving BSs, and, as a consequence, the inter-operator interference is small, both

When the operators are impatient, we can view each stage as a separate one-shot game in which reciprocity must be satisfied. Thus, each operator gives and takes an equal number of favors at each stage game. Since the operators are selfish, a favor is exchanged only if both operators experience a positive reward at that stage.

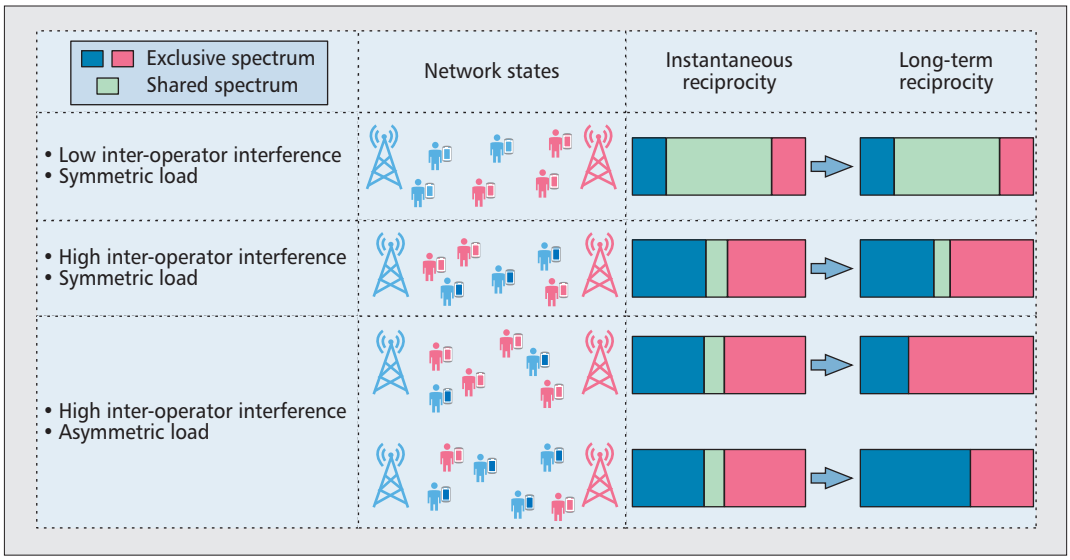


Figure 2. Visualizing the functionality of the proposed coordination protocol based on the combined short-term and long-term reciprocity over four stage games.

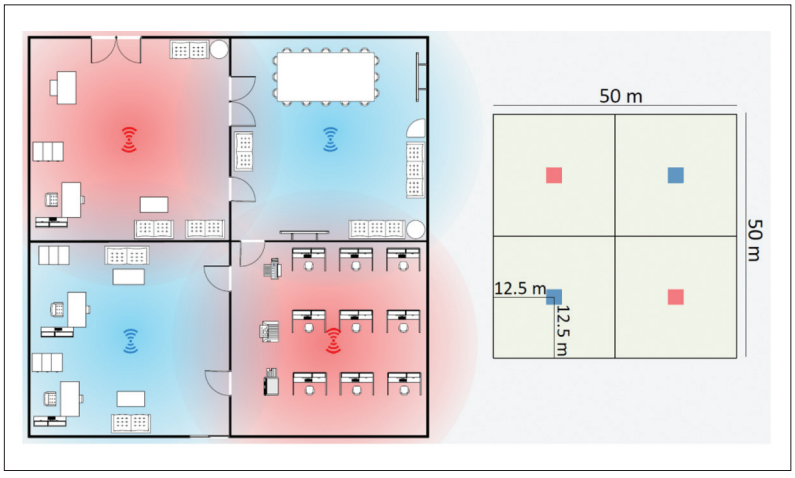


Figure 3. Indoor inter-operator deployment scenario. Different colors represent BSs of different operators.

operators would be eager to share spectrum. On the other hand, when UEs are exposed to high inter-operator interference, the operators would prefer to orthogonalize their resources. Furthermore, patient operators would be willing to empty spectrum resources when they have few or no UEs to serve, and request more spectrum resources when they really need them.

Irrespective of the spectrum sharing scenario, the one-shot game adapts spectrum sharing to the inter-operator interference conditions. The outcome of the one-shot game does not depend on the RAN load variations. This is where repeated games can come into play by allocating more spectrum to the operator with higher load, provided that this operator has been cooperative in the past. We thus consider a combined protocol where in each instance of the game operators first exchange spectrum usage favors based on short-term reciprocity to adapt to the inter-operator interference situation, and then exchange favors based on long-term reciprocity to exploit network traffic dynamics. The functionality of the proposed coordination protocol is illustrated in Fig. 2.

Numerical Illustrations

We study the UE rate improvement for operators applying spectrum sharing coordination protocols in an MR scenario. We consider an indoor deployment with two LTE small cell operators in a single-story 50 × 50 m^2 building (Fig. 3). The building has four identical rooms, and an operator's BSs are stationed in diagonal rooms. We consider downlink transmissions, a proportionally fair utility function for both operators, and a full buffer traffic model. Thus, the number of UEs represents the network load, which is generated from a Poisson distribution with possibly different means for different operators. The UEs are uniformly distributed in the service area of operators. Cell association is based on received signal power.

We consider the WINNER indoor office path loss model [12]. A bandwidth of 80 MHz is split into four CCs, with each operator owning an exclusive license for two CCs. The available power budget per CC is 20 dBm, and the aggregate external interference plus noise level per CC is –80 dBm. Each operator contributes a CC to the spectrum sharing game and reserves the other for its own exclusive usage. Fig. 4 depicts the strategy profile for operator A.

At each stage game, operators first play the one-shot game and select strategy 1 or 2 in Fig. 4. Based on the minimum rule, the outcome with the least carrier sharing is selected. For instance, the outcome of the one-shot game is 2 only if both operators propose to share a CC. When one or both operators are unwilling to share a CC, the outcome is 1. Next, the repeated game is executed. For example, if the outcome of the one-shot game is 2, and operator A asks for an exclusive favor of two CCs from operator B, which operator B grants, the result is 2c.

We evaluate the performance of the combined coordination protocol over a finite time

horizon of 4000 stage games. First we consider a scenario with equal mean network loads between the operators and low inter-operator interference. The coverage areas are non-overlapping; the UEs are served by a BS in the same room, as in Fig. 3. The mean number of UEs for each operator is 5, and the wall loss is 10 dB.

In Fig. 5, the rate distribution for the UEs of an operator is depicted. The QoS with sharing is significantly better than without. The rate improvement in the 10th percentile of user rate cumulative distribution function (CDF) is 47 percent, and in the 50th percentile is 45 percent. The full spectrum sharing outcome (2a) is most likely. This is the ideal solution in a low interference environment. The gains of the combined protocol and the one-shot scheme are virtually the same. Only a few favors are exchanged during the repeated game as the operators' loads are similar. In addition, the outcome of a fully cooperative protocol is depicted, where the operation of the two networks is jointly optimized. The full cooperation results do not differ from those of the coordination protocol.

Next, we analyze a situation with load asymmetry and high inter-operator interference. The UEs of both operators are distributed uniformly in the whole building with no internal walls. In half of the instances, the mean number of UEs is 8 for operator A and 2 for operator B, whereas in the other half, the loads are reversed.

In Fig. 6, we see a decline in UE rate in comparison to Fig. 5 due to intense inter-operator interference. One-shot sharing provides a small rate improvement. A marginal improvement of 1 percent in the 10th percentile UE rate is seen, and an increment of 8 percent in the 50th percentile. In the combined coordination protocol, operator B grants more favors than operator A when it has low load, as it can cope with fewer CCs. When it has high load, it asks for and is granted more favors. The overall performance is better than for the one-shot game. The 10th percentile UE rate is improved by 26 percent and the 50th percentile rate by 23 percent compared to no sharing. The combined coordination protocol performs close to a fully cooperative joint optimization of the networks.

In the simulation setting, UEs with low rate in no sharing are UEs in a high-load instance that are far from the serving BS and accordingly close to an opponent BS. These UEs do not benefit from the one-shot game, as they do not benefit from shared spectrum. However, the repeated game can provide more resources to these UEs, and accordingly a better rate. On the contrary, UEs that experience a high rate in no sharing are UEs close to the serving BS, and accordingly far from opponent BSs. These UEs are low-interference UEs that benefit from shared spectrum provided by the one-shot game, whereas the repeated game has little effect on their performance.

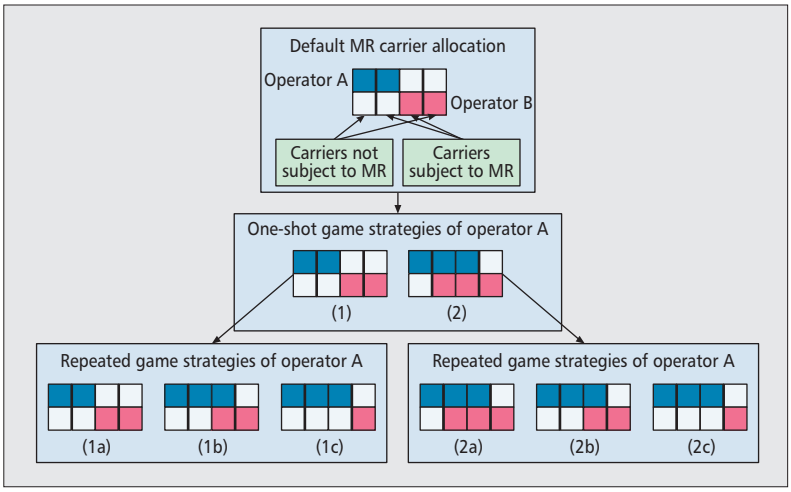


Figure 4. Strategy profile for operator A.

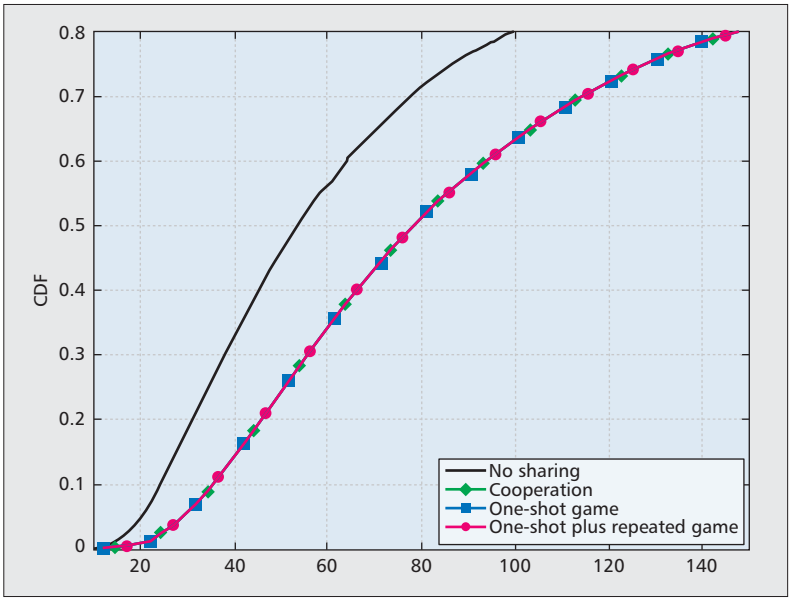


Figure 5. Rate distribution for the UEs of an operator. Equal mean network load for the two operators, low inter-operator interference.

Conclusions

The principle of allocating spectrum to mobile network operators based on a dedicated and exclusive license will persist, as a method to ensure coverage and QoS. Nevertheless, the high capacity demands generated in hotspot areas require heterogeneous network structures. Operation of small cell networks only based on dedicated licenses may not be feasible as more spectrum is needed, and new spectrum is expensive and difficult to identify. Flexible spectrum use and co-primary spectrum access is a way forward for indoor small cells. Different operators providing wireless data access in spatially separated indoor areas, or with highly directive millimeter-wave technologies, may benefit from spectrum sharing due to negligible inter-operator interference. The spectrum needs for such operators would vary in space. With overlapping coverage areas, inter-operator interference may be significant, especially for centimeter-wave and lower frequencies. Load variations and changing user locations make this interference highly variable in small cell networks, so spectrum needs of operators would vary in space and time. In order to exploit variations in spectrum needs, inter-operator spectrum sharing is a viable option. To

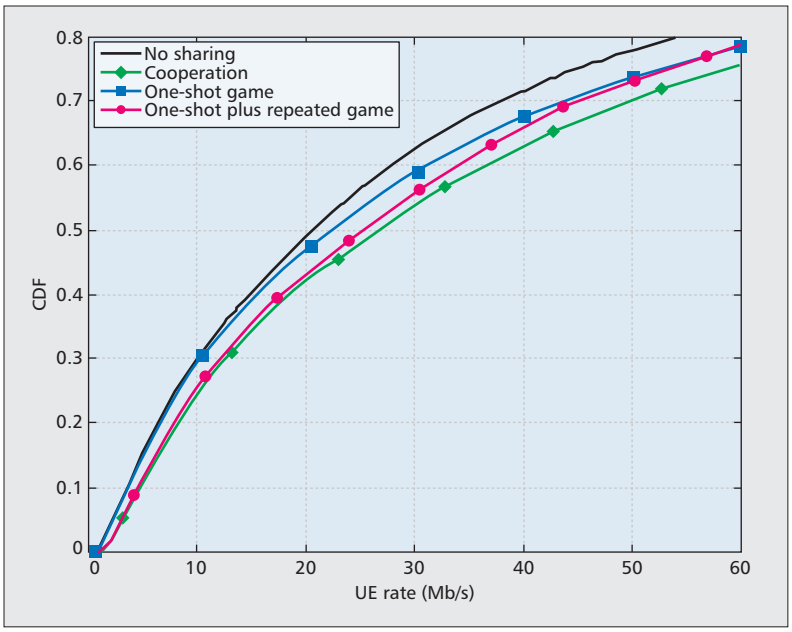


Figure 6. Rate distribution for the UEs of an operator. Unequal mean network loads for the two operators, high inter-operator interference.

realize it, we need a coordination protocol that has low implementation complexity, is transparent to the operator's revenue model, and does not require excessive information exchange among operators. Such a protocol would also enable the emergence of new types of market players (e.g., local operators). We have designed a protocol that adapts spectrum allocation to inter-operator interference situations and network traffic dynamics that are expected to be prominent in small cells. We have illustrated that in an indoor deployment scenario, two operators are both able to offer higher user rates than they could without coordination. Our results show that a rational operator, knowing that the opponent is rational and has a network with similar characteristics, has incentive to coordinate spectrum usage. Beyond improvements in service rates, there are other issues that can influence operators' incentives to coordinate, such as the fact that service quality is a key way for operators to differentiate from their competitors. The discussed protocol can be used in any spectrum reserved for mobile communication, such as in licensed shared access spectrum [13].

Acknowledgment

This work was supported in part by the European Commission in the framework of the FP7 project ITC-317669 METIS.

References

[1] A. Apostolidis *et al.*, "Intermediate Description of the Spectrum Needs and Usage Principles," METIS, doc. ICT-317669-METIS/D5.1, 2013.
[2] T. Irnich *et al.*, "Spectrum Sharing Scenarios and Resulting Technical Requirements for 5G Systems," *Proc. IEEE PIMRC*, Sept. 2013, pp. 127–32.
[3] G. Middleton *et al.*, "Inter-Operator Spectrum Sharing in a Broadband Cellular Network," *Proc. IEEE ISSSTA*, Aug. 2006, pp. 376–80.
[4] L. Anchora *et al.*, "Capacity Gains Due to Orthogonal Spectrum Sharing in Multi-Operator LTE Cellular Networks," *Proc. ISWCS*, Aug. 2012, pp. 286–90.
[5] E. A. Jorswieck *et al.*, "Spectrum Sharing Improves the Network Efficiency for Cellular Operators," *IEEE Commun. Mag.*, vol. 52, no. 3, Mar. 2014, pp. 129–36.
[6] J. E. Suris *et al.*, "Cooperative Game Theory for Distributed Spectrum Sharing," *Proc. IEEE ICC*, June 2007, pp. 5282–87.
[7] J. Huang, R. Berry, and M. Honig, "Auction-Based Spectrum Sharing," *Mobile Net. Appl.*, vol. 11, no. 3, Jun. 2006, pp. 405–18.
[8] R. Etkin, A. Parekh, and D. Tse, "Spectrum Sharing for Unlicensed Bands," *IEEE JSAC*, vol. 25, no. 3, Apr. 2007, pp. 517–28.
[9] M. Mueck *et al.*, "ETSI Reconfigurable Radio Systems: Status and Future Directions on Software Defined Radio and Cognitive Radio Standards," *IEEE Commun. Mag.*, vol. 48, no. 9, Sept. 2010, pp. 78–86.
[10] S. Hailu *et al.*, "One-Shot Games for Spectrum Sharing Among Co-Located Radio Access Networks," *Proc. IEEE ICCS*, Nov. 2014, pp. 61–66.
[11] B. Singh *et al.*, "Co-Primary Inter-Operator Spectrum Sharing over a Limited Spectrum Pool Using Repeated Games," *Proc. IEEE ICC*, June 2015, pp. 1–6.
[12] P. Kyösti *et al.*, "WINNER II Channel Models," tech. rep. IST-4-027756 WINNER II D1.1.2 V1.2, 2007.
[13] ECC, "Technical and Operational Requirements for the Operation of White Space Devices Under Geo-Location Approach," rep. 186, 2013.

Biographies

Bikramjit Singh (bikramjit.singh@aalto.fi) received his M.Sc. degree in communications engineering from Aalto University, Finland, in 2014. Currently, he is pursuing a D.Sc. (Tech.) degree in communications engineering from Aalto University in the field of spectrum sharing in 5G heterogeneous networks.

Sofonias Hailu (sofonias.hailu@aalto.fi) received his M.Sc. degree in communications engineering from Aalto University in 2014. Currently, he is pursuing a D.Sc. (Tech.) degree in communication engineering from Aalto University in the field of 5G mobility, spectrum, and radio resource management.

Konstantinos Koufos (konstantinos.koufos@aalto.fi) obtained his diploma in electrical and computer engineering from Aristotle University, Greece, and his M.Sc. and D.Sc. in radio communications from Aalto University. He is currently a post-doctoral researcher at the Department of Communications and Networking, Aalto University. His current research interests are in interference modeling and spectrum sharing.

Alexis Dowhuszko (alexis.dowhuszko@aalto.fi) received his telecommunications engineer degree from Blas Pascal University, Argentina, in 2002, and his Ph.D. degree in engineering sciences from the National University of Cordoba, Argentina, in 2010. He is currently a post-doctoral researcher at the Department of Communications and Networking, Aalto University. His research interests are in radio resource management for ultra-dense wireless networks.

Olav Tirkkonen (olav.tirkkonen@aalto.fi) received his M.Sc. and Ph.D. in theoretical physics from Helsinki University of Technology, Finland. Currently he is an associate professor in communication theory at the Department of Communications and Networking, Aalto University. His current research interests are in coding theory, multiantenna techniques, and cognitive and heterogeneous cellular systems.

Riku Jäntti (riku.jantti@aalto.fi) is an associate professor in communications engineering and head of the Department of Communications and Networking at Aalto University School of Electrical Engineering. He received his M.Sc. in electrical engineering in 1997 and D.Sc. in automation and systems technology in 2001, both from Helsinki University of Technology. His research interests are in machine type communications, cloud-based radio access networks, spectrum and co-existence management, and RF inference.

Randall Berry (rberry@eecs.northwestern.edu) received his M.S. and Ph.D. in electrical engineering and computer science from the Massachusetts Institute of Technology in 1996 and 2000, respectively. He is currently a professor in the Department of Electrical Engineering and Computer Science at Northwestern University. His research interests include wireless communications and network economics.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
COMMUNICATIONS STANDARDS SUPPLEMENT

BACKGROUND

Communications Standards enable the global marketplace to offer interoperable products and services at affordable cost. Standards Development Organizations (SDOs) bring together stake holders to develop consensus standards for use by a global industry. The importance of standards to the work and careers of communications practitioners has motivated the creation of a new publication on standards that meets the needs of a broad range of individuals including: industrial researchers, industry practitioners, business entrepreneurs, marketing managers, compliance/interoperability specialists, social scientists, regulators, intellectual property managers, and end users. This new publication will be incubated as a Communications Standards Supplement in *IEEE Communications Magazine*, which, if successful, will transition into a full-fledged new magazine. It is a platform for presenting and discussing standards-related topics in the areas of communications, networking and related disciplines. Contributions are also encouraged from relevant disciplines of computer science, information systems, management, business studies, social sciences, economics, engineering, political science, public policy, sociology, and human factors/usability.

SCOPE OF CONTRIBUTIONS

Submissions are solicited on topics related to the areas of communications and networking standards and standardization research, in at least the following topical areas:

Analysis of new topic areas for standardization, either enhancements to existing standards, or of a new area. The standards activity may be just starting or nearing completion. For example, current topics of interest include:
- 5G radio access
- Wireless LAN
- SDN
- Ethernet
- Media codecs
- Cloud computing

Tutorials on, analysis of, and comparisons of IEEE and non-IEEE standards. For example, possible topics of interest include:
- Optical transport
- Radio access
- Power line carrier

The relationship between innovation and standardization, including, but not limited to:
- Patent policies, intellectual property rights, and antitrust law
- Examples and case studies of different kinds of innovation processes, analytical models of innovation, and new innovation methods

Technology governance aspects of standards focusing on both the socio-economic impact as well as the policies that guide it. This would include, but are not limited to:
- The national, regional, and global impacts of standards on industry, society, and economies
- The processes and organizations for creation and diffusion of standards, including the roles of organizations such as IEEE and IEEE-SA
- National and international policies and regulation for standards
- Standards and developing countries

The history of standardization, including, but not limited to:
- The cultures of different SDOs
- Standards education and its impact
- Corporate standards strategies
- The impact of Open Source on standards
- The impact of technology development and convergence on standards

Research-to-Standards, including standards-oriented research, standards-related research, research on standards

Compatibility and interoperability, including testing methodologies and certification to standards

Tools and services related to any or all aspects of the standardization lifecycle

Proposals are also solicited for Feature Topic issues of the Communications Standards Supplement.

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

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

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

Spectrum and License Flexibility for 5G Networks

Adrian Kliks, Oliver Holland, Arturo Basaure, and Marja Matinmikko

ABSTRACT

Spectrum sharing is a key solution facilitating availability of the necessary spectrum for 5G wireless networks. This article addresses the problem of flexible spectrum sharing by the application of adaptive licensing among interested stakeholders. In particular, it acts as a proponent of "pluralistic licensing" and verifies it in three simulation scenarios that are of strong interest from the perspective of 5G networks. The concluding analysis offers discussion of the potential benefits offered to spectrum holders and other interested players through the application of the pluralistic licensing concept.

INTRODUCTION

Forecasts for global mobile data traffic anticipate continued strong growth [1]. While substantial technological developments are expected to improve system capabilities of fifth generation (5G) networks, additional spectrum flexibility is needed to accommodate the predicted traffic growth of mobile/wireless communications and other services.

Spectrum regulation has traditionally relied on the two extremes of exclusive use and licence-exempt access. The primary means of spectrum management thus far has been through the exclusive or "command and control" approach, which can eliminate harmful interference to licensed users to ensure reliable communications. Such approaches have proven to work very well in supporting a range of services, including ubiquitous voice connectivity in public land mobile, early generations of data services (e.g., 2G/3G/3.5G), high-quality broadcast services, and guaranteed access for critical services (military, air traffic control, emergency services, etc.).

At the other extreme, licence-exempt operation using industrial, scientific, and medical (ISM) bands has led to a rapid, and largely unforeseen, surge in wireless devices and systems including Wi-Fi, Bluetooth, and others. The ISM bands have indeed proven to be a hotbed of innovation as well as an entry point for "free" wireless communications. However, the resulting rich eco-system is built on just one caveat, that is, low transmit power on the premise of limited propagation, which hampers the scope of wireless services. For cellular networks, ISM bands have not been attractive due to the associated uncontrolled interference environment resulting in unpredictable quality of service (QoS).

In this article, we discuss the idea of flexible licensing, which will provide new spectrum opportunities for 5G systems and deliver new opportunities for spectrum holders to make additional profit gains by reusing portions of locally unused spectrum.This article discusses the various concepts of spectrum sharing presented from the perspective of their potential application in 5G networks. We then continue with the discussion of the benefits of the adoption of the "pluralistic licensing" (PL) concept contrasted with other spectrum sharing approaches, such as the two already mentioned extremes, and the novel concept of licensed shared access [2, 3]. The three conducted simulations are described showing the rational profits that can be achieved by spectrum holders. Finally, concluding remarks are provided.

SPECTRUM SHARING STRATEGIES FOR 5G NETWORKS

The density and variety of wireless services and users have dramatically increased over the past decade. The two extreme regulatory approaches of exclusive use and license-exempt access can no longer offer appropriate characteristics to satisfy future demand for wireless services, which need to balance interference tolerance, service prioritization, cost, and market suitability. 5G networks require a significant amount of new spectrum to respond to growing traffic demand, and spectrum sharing through flexible spectrum licensing is a key means to accomplish this. As we argue in this article, flexible licensing can both ensure the necessary QoS for primary and secondary spectrum users, and allow the necessary degree of sharing to alleviate future 5G capacity demand, reflected in ways such as an increase in the realized monetary value of the spectrum.

Adrian Kliks is with Poznan University of Technology.

Oliver Holland is with King's College London.

Arturo Basaure is with Aalto University.

Marja Matinmikko is with VTT Technical Research Centre of Finland.

0163-6804/15/$25.00 © 2015 IEEE

Light Licensing and Licensed Shared Access

Flexible spectrum licensing could better respond to future needs by allowing spectrum that is not used by one network in certain locations to be opportunistically used to the benefit of other 5G operators (e.g., for a fee). Two initial "compromise" regulatory solutions that are at least partially moving toward such flexibility are light licensing [4] and licenced shared access (LSA) [5]. The light licensing concept can ease the burden of coordination, registration, licensing, and interference consideration when making new frequency assignments, or in coordinating sharing between primary and secondary users. It usually implies setting license fees that just cover administrative costs, and is used in cases where there is a need to coordinate with an incumbent user or in a private commons approach where individuals and licensed users set the conditions for license-exempt access. The light licensing model tends to be used mainly for systems with no or limited interference potential, which could be further authorized by a simple automated check-in to an online light licensing tool to manage the interference spatially. Therefore, this approach is not at all suitable for scalable and longer-range high-transmission-power services.

The LSA approach is a relatively new industry-driven concept where additional licensed users are authorized to access incumbent (primary) users' spare spectrum within their licensed bands but under tight controls to prevent any disruption. It was originally intended to support business cases for the mobile broadband, where it is both economically and technically feasible. It is notable that the EC's Radio Spectrum Policy Group (RSPG) has acknowledged LSA in [6], asserting that indeed an LSA licensee might be granted the right to utilize under-used spectrum without interfering with the incumbent user. The objective of LSA is to grant additional spectrum rights of use in specific bands on a shared basis, allowing predictable QoS for all rights holders [5].

Through the aforementioned efforts and other initiatives, regulators across the globe have started to promote spectrum sharing [7, 8]. They have acknowledged that any such expanded sharing would require new regulatory paradigms, such as spectrum sharing contracts and shared spectrum access rights, to ensure the legal certainty and rules, as well as the obligations of the interacting spectrum users.

Pluralistic Licensing: The Core Concept

The PL concept was proposed by us in [9] as a novel approach, in line with a wide range of spectrum sharing contracts and shared spectrum access rights as discussed above. PL is an innovative means to improve spectrum licensing, which is fair to both primary and secondary users and takes into account the requirements of both parties. The concept is described as "the award of licenses under the assumption that opportunistic secondary spectrum access will be allowed, and that interference may be caused to the primary with parameters and rules that are known to the primary at the point of obtaining the license" [9]. The general assumption is that the primary will choose from a range of offered PLs, each with a different fee structure, and each specifying alternative opportunistic access rules that can be mapped to associated interference characteristics. The locus of control, therefore, remains firmly with the primary, whereby the primary might trade off the form and degree of opportunistic access for a reduced licensing fee or another incentive.

Under the PL concept, primary users who obtain the license, which might be on a very short-term or longer-term basis or even geographical (e.g., per-transmitter), are allowed to access the spectrum at will. A coordination mechanism among primary assignments would nevertheless be needed in cases where there are multiple primaries coexisting. Secondary users must use a "cognitive" mechanism to access the band, whereby the detail of the cognitive mechanism (the use and configuration of a spectrum sensing approach and/or a geolocation database, etc.), as well as its radio characteristics, depends on the context within which the band is chosen to operate. This context might include the expected types of primary services(s) in the band (perhaps also expected secondary services(s)), an assessment of an appropriate "burden" on the primaries in terms of acceptance of a slightly higher probability or net amount of interference while still achieving adequate performance, the degree to which the primary and secondary should negotiate, or even in some foreseeable cases the degree to which the primary should be expected to take proactive measures such as the transmission of beacons. Of course, such a context defines the extent to which the secondary must avoid interfering with the primary, and hence the associated rules on the secondary. Crucially, in this sense, PL can be the practical form of implementing the *spectrum sharing contracts* already envisaged by regulators [8].

Benefits Achieved through Flexible Licensing

There are numerous benefits of the concept of PL, which are directly tangible to 5G networks and other players in the market. First, particularly in green-field scenarios, there is no need for the secondary systems to cope with the inefficiencies of legacy primary systems, as can be an issue in other spectrum sharing realms such as TV white space. In obtaining the license, the "primary" 5G operators will implicitly accept (and even decide); hence, the rules of the band will be designed and manufactured with better technical capabilities to cope with those rules—in return for an incentive such as a reduced license fee. For instance, the operators holding such licenses might deploy systems that can better reject adjacent channel interference, or enhance their adaptive rate and error correction mechanisms when the opportunistic secondary spectrum access might imply a higher probability of an interference limit being violated or a higher variance in the experienced interference to the other 5G network. Furthermore, such a concept might be applied to other primary services, with the 5G network effectively being the opportunistic spectrum user of that service's spectrum.

There are numerous benefits of the concept of PL, which are directly tangible to 5G networks and other players in the market. First, particularly in green-field scenarios, there is no need for the secondary systems to cope with the inefficiencies of legacy primary systems, as can be an issue in other spectrum sharing realms such as TV white space.

A further, more general benefit of the concept is that it satisfies the need for spectrum sharing between more established users and incumbents, generally giving the 5G network guaranteed spectrum access with a given QoS, while still allowing free access when/where the spectrum is not used by that network.

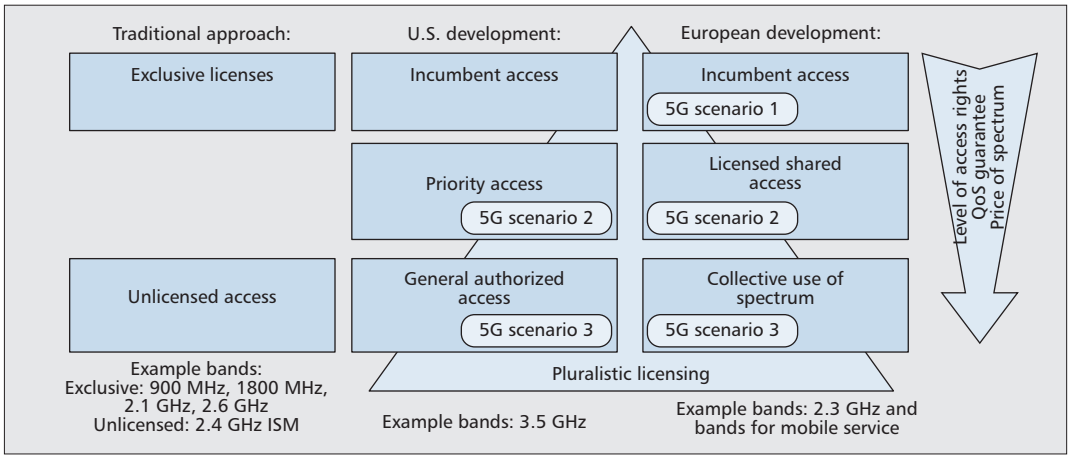


Figure 1. Pluralistic licensing in the context of spectrum sharing approaches for 5G.

A further, more general, benefit of the concept is that it satisfies the need for spectrum sharing between more established users and incumbents, generally giving the 5G network guaranteed spectrum access with a given QoS, while still allowing free access when/where the spectrum is not used by that network. This concept can be used to strike a good balance between operators needing to pay more for quality in some locations, and perhaps less in others in return for allowing forms of opportunistic access. Indeed, it is even shown later in this article that this context might even be used to very significantly *increase* profit for the operator.

An additional benefit is that the concept very likely implies significantly improved performance, such as in terms of spectrum usage efficiency. This is very compelling for the 5G operator in scenarios in which the operator might have an agreement with particular systems, and be able to extract revenue for the opportunistic access and increased efficiency of its spectrum usage. A further benefit of the concept is high scalability to progressive deployment in more spectrum bands. This realization is very much in line with 5G networks, which are increasingly likely to be designed and built on large chunks of highly distributed spectrum.

Application to Opportunistic Access and Spectrum Sharing in 5G Networks

PL is seen as applicable to almost any spectrum licensing scenario, and indeed is viewed by the authors as one possible panacea to the problem of licensing while allowing spectrum sharing in an agreeable and fair way for all concerned. For application to 5G networks, the primary (network operator) will generally have a good understanding of the effects each possible form of secondary access is likely to have. This will, of course, depend on the type of secondary system and associated characteristics such as mobility and transmission patterns of the radio interface (MAC/PHY, bandwidth, and frequency), as well as possibly higher-layer characteristics. It will intrinsically also be linked to the configuration of the network that the primary is deploying and its associated requirements. The latter can be controlled by the primary network operator, but the former must be predicted and mitigated through the choice of PL and associated characteristics.

Aside from such technical considerations, economic considerations can also be expanded upon. The traditional exclusive licenses have strengthened the dominating role of the big operators, as only they can afford to buy licenses from costly auctions. The recent technology development in cellular networks is prepared for spectrum sharing including many supporting features, such as self-organizing networks (SONs), carrier aggregation, and hierarchical cell layers, which can already share spectrum. By allowing local temporary licenses with agreed conditions through the PL approach, operators may adjust their spectrum assets more dynamically to respond to actual needs, and new entrants might get access to spectrum resources, facilitating the more optimal realization of resources right up to the level of dynamic aggregation of tailored spectrum opportunities matching the QoS demands of higher-layer traffic.

The evolution of the spectrum sharing regulatory framework is focusing on three general levels of access rights: primary access, secondary access, and collective use [7, 8]. While in the past cellular networks were solely deployed on exclusively licensed bands, 5G networks are expected to operate on different types of spectrum bands to meet the growing demand. Following the generic spectrum sharing framework, the PL approach could be applied to 5G networks in several scenarios, as depicted in Fig. 1, allowing the operators to acquire different types of spectrum assets according to their preferences and needs.

In the first scenario, the 5G networks could obtain primary access rights to a given spectrum band similar to the traditional exclusive licensing, but it could also admit secondary access/licensed shared access and general authorized access/collective spectrum use rights to other users to access its band with predetermined conditions and rules, and benefit from it. In fact, the regulator could endorse this sharing approach by collecting lower spectrum fees from 5G network operators that allow access to their bands. Moreover, by using the PL concept to allow secondary access/licensed shared access, a 5G operator could collect fees from secondary users. The first scenario could be applicable to the potential new

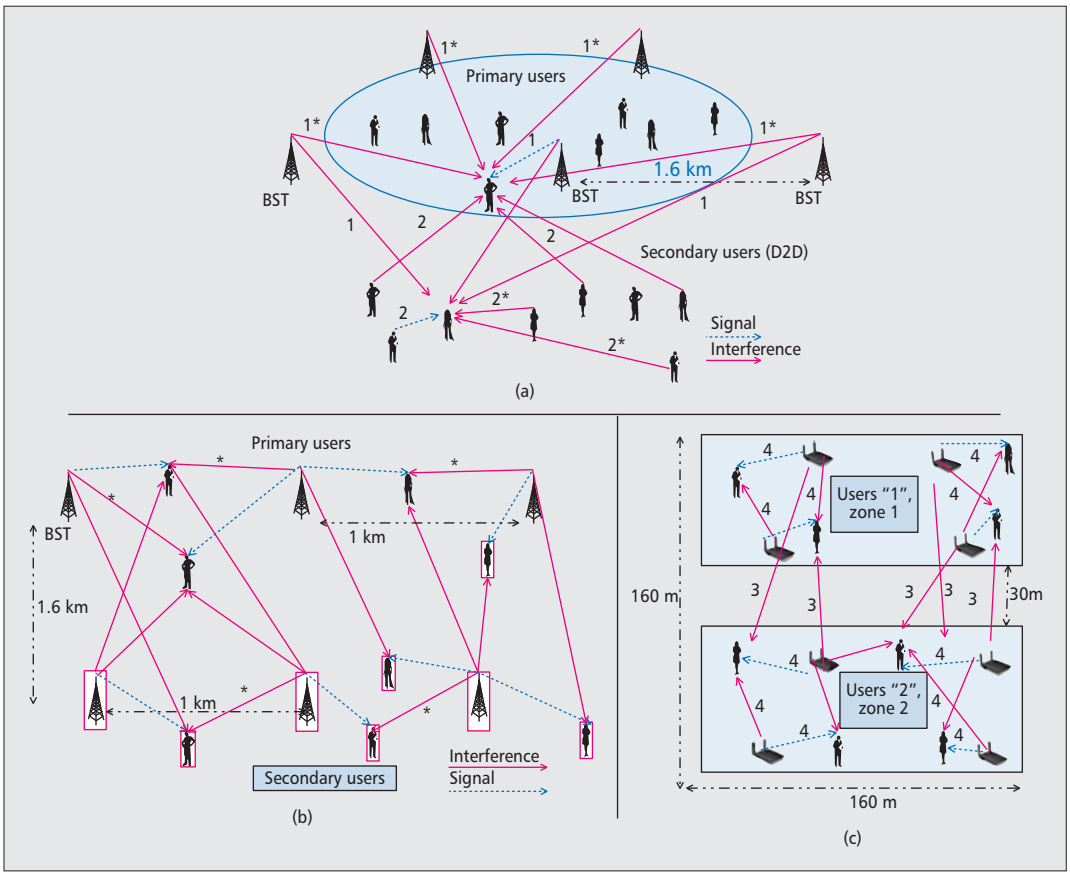


Figure 2. Three considered scenarios: a) FDD downlink transmission with D2D communication; the numbers above the arrows define the type of path loss models used: 1. COST Hata model, 2. ITU-R P.1411 end; b) cross-operator spectrum sharing; c) indoor-indoor co-primary sharing with the following path loss model: 3. In-building propagation model of [11, Eq. 5] with wall penetration loss: 4. In building propagation model path loss model.

By allowing local temporary licenses with agreed conditions through the PL approach, operators might adjust their spectrum assets more dynamically to respond to the actual needs, and new entrants might get access to spectrum resources, facilitating the more-optimal realization of resources.

spectrum bands with primary allocation to mobile that would be cleared from other use.

In the second scenario, the 5G network would get secondary access/licensed shared access rights to a given spectrum band using the PL approach, which would determine the rules and conditions that guarantee the primary users remain free from harmful interference but at the same time offer sufficient QoS for the 5G network. In this scenario, the 5G networks could gain access to spectrum resources with reduced costs compared to the traditional exclusive licenses by acquiring a local and temporary license depending on its needs. It could gain access to new spectrum bands that with traditional regulatory approach are not available as they are primarily allocated to other use but whose actual occupancy may remain low. Offering PLs with fair conditions to both primary access and secondary/licensed shared access users would open up a considerable amount of new spectrum for 5G networks with QoS conditions resembling exclusive licensing but with lower costs.

In the third scenario, the next generation network could exploit general authorized access/collective spectrum use rights to access spectrum bands that are allowed to be used by multiple users simultaneously according to a set of predefined rules with little or no cost. While these bands would not necessarily offer high QoS over a large geographical area, they could be used for short-range mobile data offloading as is currently done using the unlicensed access mode with Wi-Fi.

Simulation Scenarios for Spectrum Sharing Incentivization

In order to illustrate the benefits of the application of spectrum sharing strategies based on the PL approach, a set of simulation experiments has been carried out with the use of agent-based modeling. Three scenarios have been identified that are of high interest from the perspective of the spectrum holder in the context of 5G spectrum sharing. In the first scenario, short-range transmissions between two devices (e.g., the device to device, D2D, case) are considered, as realized through PL in the same frequency band as the nearby cellular network using a database or spectrum sensing.

In the second scenario, the coexistence of two networks in close geographical proximity is analyzed assuming one of the operators would like to share the spectrum assigned originally to another operator (primary user), applying the concept of flexible PL.

The final simulation scenario provides an

Assumptions			
Scenario (primary/secondary) Parameters	Scenario 1	Scenario 2	Scenario 3
Cellular BS spacing	1.6 km	3.2 km	NA
Secondary leasing distance	100, 200, and 300 m	100 m	NA
Separation between coverage areas	700 m	400 m	30 m
BS transmission center frequency	2.6 GHz	2.6 GHz	60 GHz
Primary BS transmission power	From 40 to 52 dBm	From 43 to 49 dBm	From 18 to 24 dBm
Secondary BS transmission power	NA	from 43 to 49 dBm	from 18 to 24 dBm
Primary terminals transmission power	21 dBm	21 dBm	21 dBm
Secondary terminals transmission power	20 dBm	21 dBm	21 dBm
Antenna gain	BST (10 dB)	BST (10 dB)	2.1 dB
Path loss model from base station to terminal	Cost 231 Hata model	Cost 231 Hata model	Anderson and Rappaport [11, Eq. 5]
Path loss model from terminal to terminal	ITU-R P.1411 end-to-end model	NA	NA
Noise power	–105 dBm or 0.032 pW (in 8 MHz channel)	–105 dBm or 0.032 pW	–105 dBm or 0.032 pW
Shadowing standard deviation	6	6	6
BS effective height, primary case	30 m	30 m	1.5 m
BS effective height, secondary case	NA	30 m	1.5 m
Primary node effective height	1.5 m	1.5 m	1.5 m
Secondary node effective height	NA	1.5 m	1.5 m

Table 1. System setup for the considered simulation scenarios.

indoor-indoor analysis, where two sets of users located inside nearby buildings operate in the same frequency ranges, causing potential interference. All these scenarios are presented in graphic form in Fig. 2, whereas their corresponding configuration details are summarized in Table 1.

Scenario 1: PL Enhancing D2D Communications

In the first case, we consider the coexistence of a five-cell-wide section of a cellular network (treated as the primary) with direct D2D transmissions realized in the relatively small region outside the coverage area of the primary (please also see [10]). Two approaches for granting spectrum access for secondary users are considered: first, where secondary users query dedicated databases asking for transmit permission, and second, where spectrum sensing is applied. In order to minimize the number of queries sent by the secondary (D2D) devices, each granted user is obliged to send new requests to the database only when it leaves the so-called leasing region. It is assumed that the overall environmental and system parameters will not differ strongly within the small region; thus, the replies to the queries sent from any location inside that region will be almost the same. Referring to Table 1, three radii of leasing region have been considered: 100 m, 200 m, and 300 m. Moreover, in order to assess the performance of spectrum sensing, one additional leasing region of size 10 m was also applied.

In order to present the potential benefits of the PL, we analyze the averaged SINRs observed by the primary and secondary users as a function of transmit power and size of the leasing region. The achieved results are presented in Fig. 3, divided into four parts. In Fig. 3a the cumulative distribution function of the SINR observed by the primary users inside the coverage area against different primary user transmission powers is presented. In Fig. 3b, the average SINRs observed by the primary and secondary users are plotted as a function of primary base station transmit power for different radii of the leasing region. The increase of the leasing radius results in SINR degradation of both types of networks

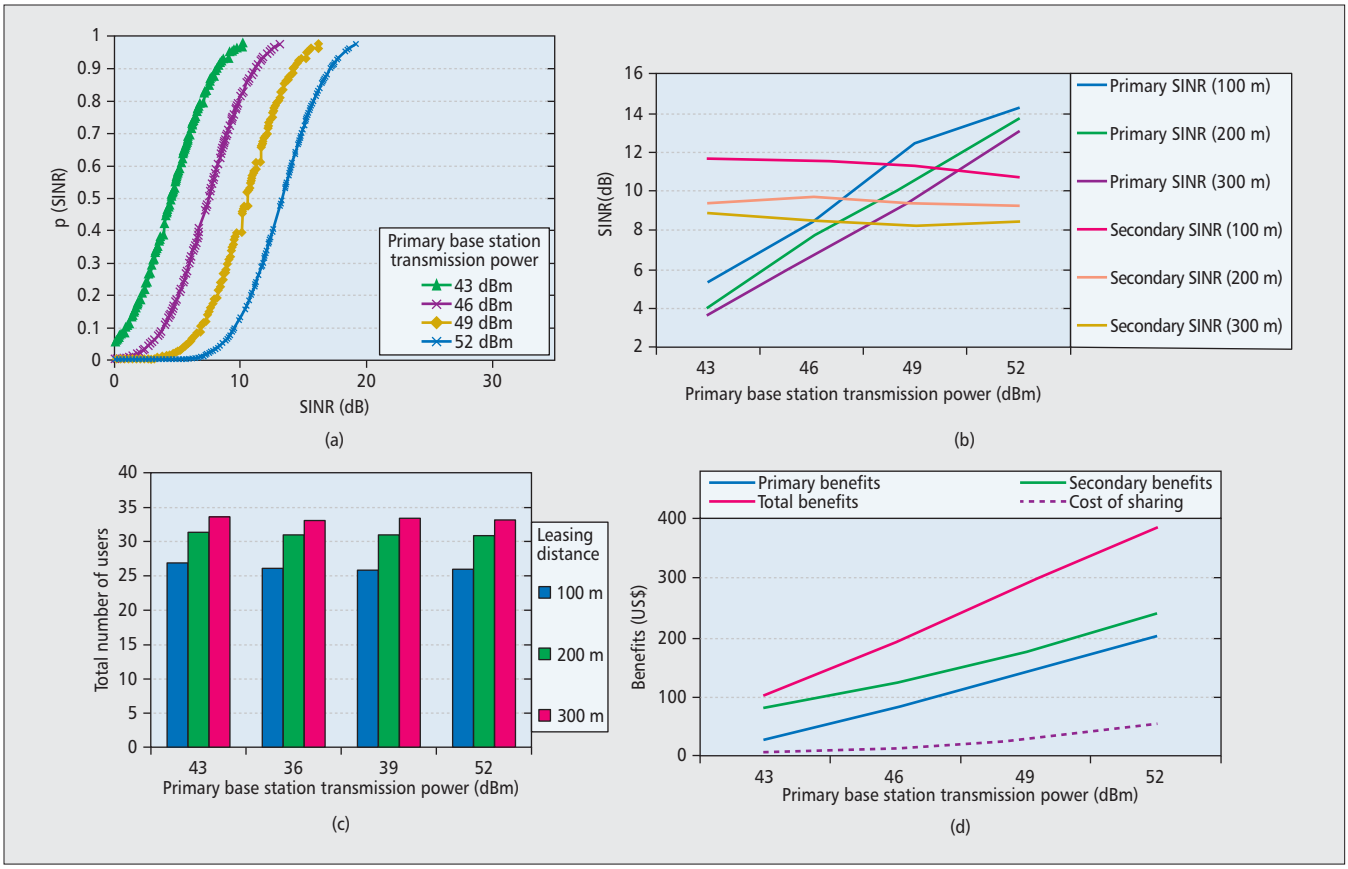


Figure 3. Results achieved for the first scenario: a) cumulative distribution function of the average SINR; b) averaged SINR observed by the primary and secondary users as the function of the leasing distance and transmit power; c) average number of supported users; d) financial benefits.

since the lower the frequency of sending queries by the secondary users, the higher interference generated to the network. Moreover, increasing transmission power leads to a very significant increase in SINR for the secondary and a slight degradation for the primary, meaning that the average link performance among the primary/secondary users is improved. In Fig. 3c, the average number of users that can be served is presented: more users can be served when the leasing distance is increased through the flexibility allowed by PL. Interestingly, this number does not depend on the maximum transmit power. It should be noted that the transmit power of the secondary D2D users is constant (20 dBm). In this case, the amount of interference produced by the secondary users is increased by adjusting the "leasing distance" (Fig. 3b, 3c), or the distance secondary users are able to move before they should request access from the spectrum database again. As shown in Fig. 3, if the distance is higher, the number of secondary users is increased at primary SINR expenses. Finally, Fig. 3d gives the profit achieved by the primary and secondary users, as well as the total profit taken from the spectrum, based on the same assumptions of traffic utility functions used in [10]. It is clear here that an increase in profit can be achieved for the spectrum as a whole by allowing an increase in the secondary transmission power through PL, whereby there is a minor impact on the primary's profit that can easily be compensated by the secondary with the secondary still making a good profit. Moreover, it is noted here that both the primary and secondary users challenge "real-time" utility functions, underpinning a high level of reliability for the D2D and cellular deployments sharing the spectrum.

Scenario 2: Coexistence of Two Cellular Networks

In the second scenario, the coexistence of two separate cellular networks located in close geographical proximity is simulated, where one of the operators (secondary user) would like to share the spectrum assigned originally to another operator (primary user), applying the concept of flexible PL (Fig. 2). Based on the setup presented in the third column of Table 1, we concentrate on measurements of the observed average SINR as a function of the maximum allowed transmit power. Downlink transmission with the FDD scheme has been simulated. Based on observed results shown in Fig. 4, one can state that there is a slight performance degradation as secondary base stations increase their transmit power. Furthermore, analogous to the previous scenario, the application of the PL concept generates a significant profit increase for the spectrum while providing appropriate compensation to the primary for the minor effect it experiences due to the increased sharing.

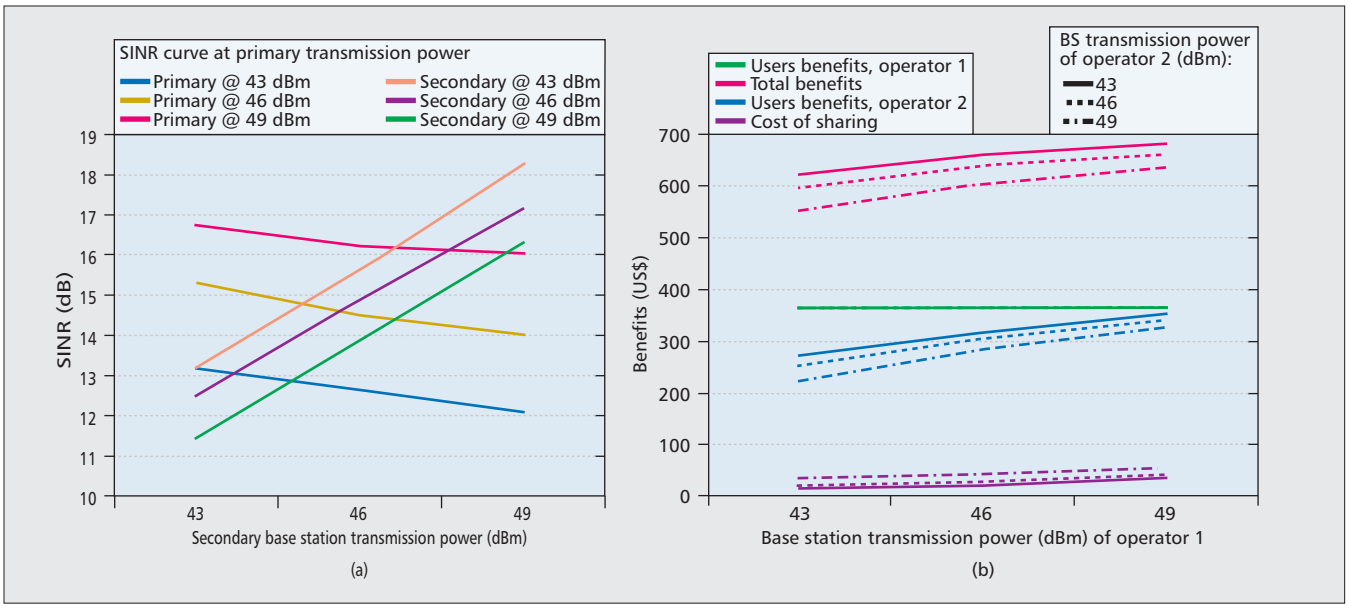


Figure 4. Cellular-cellular sharing scenario: a) average SINR of the primary and secondary users as the function of transmit powers of primary and secondary base stations; b) profits of the two sharing operators, individually and combined.

Scenario 3: Indoor-Indoor Coverage Improvement

In this final scenario, two disjoint sets of users are located inside nearby buildings (Fig. 2), connected to indoor access points (APs) hosted in each of the buildings. They share the same frequency band; hence, they can cause interference to each other. Reflecting a 5G scenario, this frequency band is assumed to be at 60 GHz. These two sets of users might be seen as co-primary users, or implementing a primary-secondary scenario with one set of users in one building being the primary, and the other set in the other building the secondary. Under our results, it is best if the sets of users are seen as co-primary; however, the simulations can also be used to infer what would happen if one set of users were the primary and the other secondary.

The simulation parameters assumed for this case are summarized in the last column in Table 1. The two-strip layered model has been applied to the buildings, and a detailed path loss model has been selected in order to consider wall attenuation [11, Eq. 5]. The goal of the conducted simulations has been to verify the influence of the maximum transmission powers used in each building on the observed averaged SINR values of the sets of users, these maximum transmission powers being variable through the flexibility brought about in a PL scenario. The corresponding results are presented in Fig. 5. It is clear that the average SINR experienced among the sets of users increases significantly if the allowed maximum transmission power is increased for either or both of the sets of users.

Rational Benefits Achieved by a Spectrum Holder

Finally, we briefly identify the benefits that can be gained by spectrum holders. Practical utilization of the PL concept allows definition of flexible pricing strategies that can be applied. Taking into account various types of traffic, it can be argued that as in one case the interference induced by the other users will potentially lead for high QoS degradation, in another scenario such interference power can easily be tolerated. Such an observation is particularly important in the context of 5G networks, where delivering various services (of different QoS levels and associated guarantees) to the user is envisaged. Thus, the primary user (e.g., network operator) can accept more interference from the other interested player (e.g., another operator, non-first-priority end users) at the price of an increased fee paid by that player. Higher interference means, in fact, a greater leasing region or higher transmit power, thus higher throughput that will be served and managed by the licensee. On the contrary, lower fees can be offered to such players that will not be allowed to transmit with maximum power.

The analysis of the results achieved in the three simulation scenarios has proven that a plethora of interesting variants for flexible spectrum sharing, and in consequence PL, can be applied. This will be attractive for both current spectrum holders and any other player interested in sharing the spectrum. Thus, the LSA approach together with the complementary PL concept can be treated as key regulatory enablers for better utilization of resources in 5G networks and spectrum availability enhancement through sharing.

Conclusions

It can be foreseen that with the introduction of new, often technically challenging services to the end user, the need for additional spectrum will significantly increase in the very near future. This has led to the conclusion that the current static spectrum management solutions will no longer be applicable, and a new vision for 5G spectrum is required. Realization of the adaptive spectrum sharing concept will definitely pave the way for more efficient utilization of spectrum

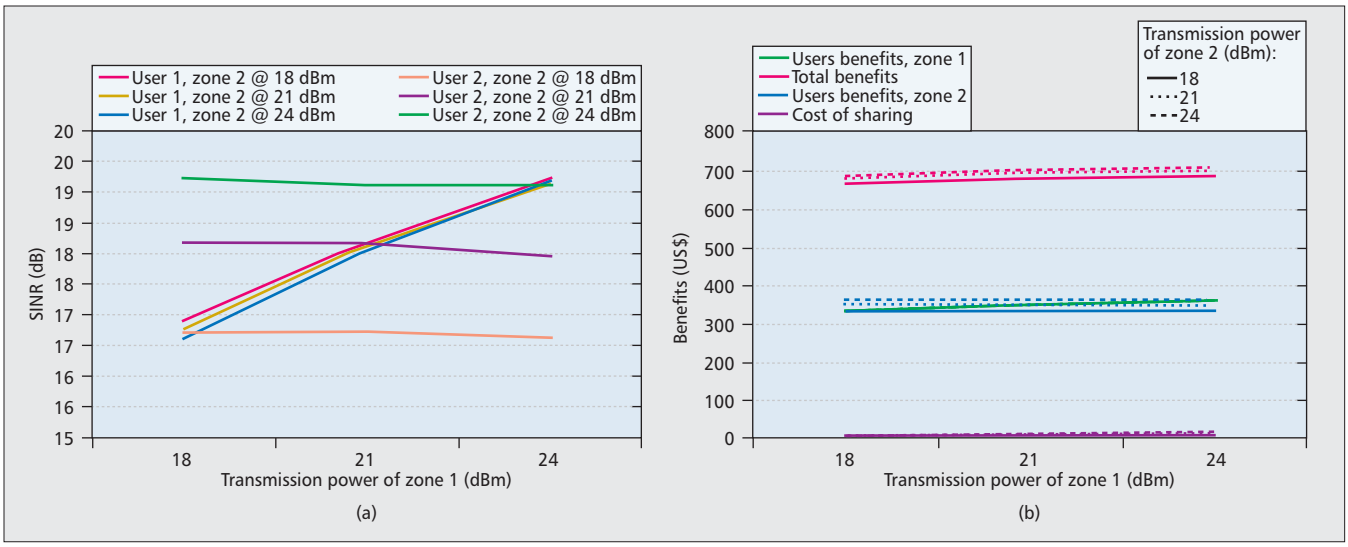


Figure 5. a) Achieved average SINRs for the sets of users in Scenario 3, against varied transmission power; and b) profits of the two sharing operators, individually and combined.

resources. Licensed shared access seems to be the first practically available solution, but much more can be beneficial from the application of the flexible pluralistic licensing concept, which at the same time delivers to the spectrum holder a new tool for revenue increase. However, such solutions have to be supported by the appropriate regulatory decisions made at the global level, which will open the doors for new definition of spectrum usage in the context of 5G networks.

Acknowledgments

The work of the first author has been funded by the EU 7th Framework Programme project NEWCOM# (contract no. 318306), and by the Polish Ministry of Science and Higher Education cofinancing this project. The second author acknowledges support of the EU 7th Framework Programme ICT-SOLDER Project (contract no. 619687), the ICT-ACROPOLIS Network of Excellence (contract no. 257626), COST Action IC0905 "TERRA," 5G-NORMA, and again NEWCOM#. The third author acknowledges funding from the EECRT project of Aalto University.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019," white paper, Feb. 2015.
[2] M. Mustonen *et al.*, "Cellular Architecture Enhancement for Supporting the European Licensed Shared Access Concept," *IEEE Wireless Commun.*, vol. 21, no. 3, June 2014, pp. 37–43.
[3] P. Ahokangas *et al.*, "Business Scenarios for Incumbent Spectrum Users in Licensed Shared Access (LSA)," *9th Int'l. Conf. Cognitive Radio Oriented Wireless Networks and Communications*, 2014, 2–4 June 2014, pp. 407–12.
[4] CEPT, "Light Licensing, License-Exempt and Commons," ECC Rep. 132, June 2009.
[5] RSPG, "RSPG Opinion on Licensed Shared Access," RSPG13-538, Euro. Commission, Nov 2013.
[6] RSPG, "RSPG11-392 Report on Collective Use of Spectrum (CUS) and Other Spectrum Sharing Approaches," Nov. 2011; http://rspg.groups.eu.int/rspg_opinions/index_en.htm, accessed May 2012.
[7] President's Council of Advisors on Science and Technology, "Report to the President Realizing the Full Potential of Government-Held Spectrum to Spur Economic growth," July 2012.
[8] Euro. Commission, "Promoting the Shared Use of Radio Spectrum Resources in the Internal Market," *EC Communication COM(2012) 478*, Sept. 2012.
[9] O. Holland *et al.*, "Pluralistic Licensing," *IEEE DySPAN 2012*, Bellevue, WA, Oct. 2012.
[10] A. Basaure and O. Holland, "Sharing Incentivization through Flexible Spectrum Licensing," *IEEE 9th Int'l. Conf. Cognitive Radio Oriented Wireless Networks and Communications*, June 2–4, 2014, pp. 401–06.
[11] C. R. Anderson and T. S. Rappaport, "In-Building Wideband Partition Loss Measurements at 2.5 and 60 GHz," *IEEE Trans. Wireless Commun.*, vol. 3, No. 3, May 2004, pp. 922–28.

Biographies

Adrian Kliks (akliks@et.put.poznan.pl) received with honors his M.Sc. and Ph.D. degrees in telecommunications from Poznan University of Technology (PUT), Poland, in 2005 and 2011, respectively, and currently is employed as an assistant professor at the Chair of Wireless Communications, PUT. He has been involved in various industrial and international research projects. His scientific interests include advanced multicarrier communications, waveform design, small cells in heterogeneous networks, WiFi offloading, and cognitive radio.

Oliver Holland (oliver.holland@kcl.ac.uk) received his B.Sc. degree with first class honors from Cardiff University and his Ph.D. degree from King's College London. He is extremely active in spectrum sharing related work. He recently led the ICT-ACROPOLIS Network of Excellence on spectrum coexistence technologies, and is leading a major trial of TV white space technologies within the Ofcom TV White Spaces Pilot. Among other activities, he leads two IEEE Standards Working Groups on spectrum sharing. He has co-authored more than 130 papers, which have been cited more than 850 times.

Arturo Basaure (arturo.basaure@aalto.fi) received his M.Sc. in telecommunications engineering from Helsinki University of Technology, Finland, in 2005. He has worked as an IT consultant and in the financial industry. Currently, he is a doctoral candidate at Aalto University, Finland, involved in the EECRT and EMERGENT research projects, and his main research area is the regulation of mobile telecommunications.

Marja Matinmikko is a senior scientist at VTT Technical Research Centre of Finland. She received her M.Sc. degree in industrial engineering and management, and her Dr.Sc. degree in telecommunication engineering from the University of Oulu. She is the coordinator of the Finnish project consortium on Cognitive Radio Trial Environments (CORE). Her current research interests include technical, trialing, regulatory, and business aspects of spectrum sharing for mobile communications.

Broadcast Television Spectrum Incentive Auctions in the U.S.: Trends, Challenges, and Opportunities

David Gómez-Barquero and M. Winston Caldwell

ABSTRACT

This article presents an overview of the upcoming television broadcast spectrum incentive auction in the U.S., which will be the first ever attempted worldwide, and discusses the main business, regulatory, and technical challenges of a successful incentive auction. The process combines two separate but linked auctions: a reverse auction, which will identify the prices at which broadcasters are willing to relinquish their spectrum; and a forward auction, which will determine the price mobile network operators are willing to pay to acquire the new frequencies. The two auctions will determine the buyers and sellers and also the amount of spectrum to be cleared in the 600 MHz band after reorganizing the television stations that remain on air. This process is known as repacking and will create contiguous blocks of cleared spectrum at the high frequency side of the UHF band for mobile use. The article also reviews the potential plans for the 600 MHz band and discusses the opportunities that could bring about the new digital terrestrial television standard known as "ATSC 3.0."

INTRODUCTION

The radio frequency (RF) spectrum is a finite natural resource with considerable economic and social importance. The ultra-high frequency (UHF) band from 470 to 862 MHz has traditionally been used for terrestrial television broadcasting [1]. The International Telecommunication Union (ITU) allocated the upper part of the terrestrial broadcasting UHF band to international mobile telecommunications (IMT) technologies during the World Radiocommunications Conferences (WRC) of 2007. This band ranges from 790 to 862 MHz in Region 1 (800 MHz band), and from 698 to 790 MHz in Region 2 and Region 3 (700 MHz band), see Fig. 1.

The term *digital dividend* was introduced to name the television spectrum released during the switch-over from analogue to digital terrestrial television (DTT). Initially, the number of available broadcast TV channels in Europe and the U.S. was 49 and 55 channels, respectively, which were reduced down to 40 and 37 after the first digital dividend. After the second digital band in the 700 MHz band, only 28 will be left in Europe, whereas in the U.S. it will depend on the outcome of the incentive auction.

Before the analogue TV switch-off took place, most spectrum regulators worldwide auctioned and awarded the digital dividend band to fourth-generation (4G) Long Term Evolution (LTE) mobile networks together with other frequency bands (e.g. 1.8 or 2.6 GHz). The U.S. was the first country to deploy nationwide 4G LTE networks in the digital dividend band. The deployments in Europe were initially hampered due to the lack of harmonized spectrum in the region, delays in the analogue TV switch-off, and the popularity of terrestrial television in some countries where DTT is the main television distribution platform. Despite the issues of the deployment of 4G networks in the digital dividend band, many networks are currently on air or are being deployed, including in the Asia-Pacific region. Discussions have turned to the feasibility of a second digital dividend in the terrestrial broadcasting UHF band.

At the WRC-2012, it was agreed to allow the introduction of mobile broadband services in the 700 MHz band in ITU Regions 1 and 3, to be effective after the upcoming WRC-2015. National administrations in those countries will have the possibility to enable a second digital dividend in that band. Some European countries such as Finland, Germany, Sweden, and the UK have already announced their intentions to release these frequencies for 4G LTE (some countries as early as 2017). Moreover, the European Commission (EC) is considering reallocating the 700 MHz band around 2020 [2]. The European Conference of Postal and Telecommunications Administrations (CEPT) has proposed that the established terrestrial broadcasting UHF spectrum below 700 MHz would continue to be used for terrestrial broadcasting and would remain in place until at least 2030.

David Gómez-Barquero is with Universitat Politecnica de Valencia.

M. Winston Caldwell is with 21st Century Fox's Fox Networks Group.

0163-6804/15/$25.00 © 2015 IEEE

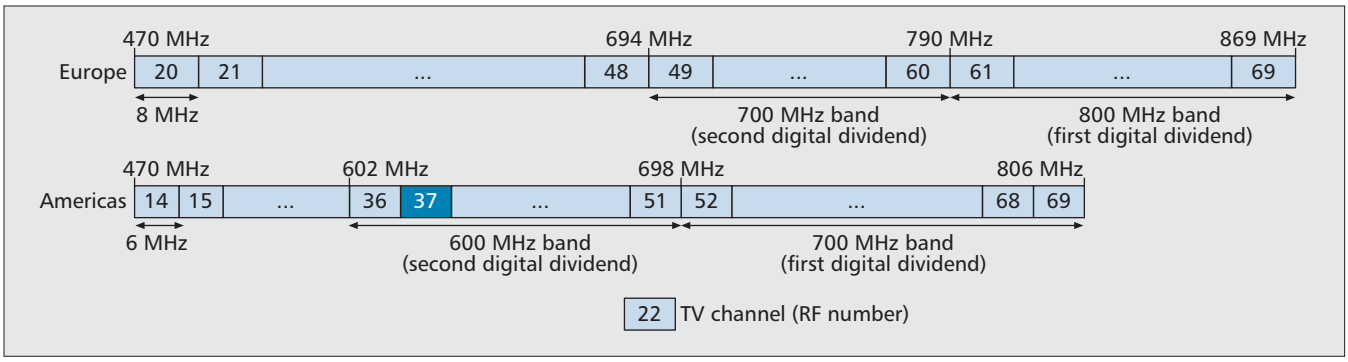


Figure 1. Frequency allocations in the UHF band for digital terrestrial broadcasting in ITU Region 1 (Europe, Middle East and Africa, and Russia) and Region 2 (Americas).

The 700 MHz band is already being used by mobile networks in the U.S. as the first digital dividend band. The U.S. is taking a further step by proposing the 600 MHz band as a second digital dividend. The Federal Communications Commission (FCC) has introduced a novel market-based spectrum auction scheme known as the *broadcast television spectrum incentive auction* [4], in which broadcasters may voluntarily relinquish their spectrum licenses in exchange for a share of the auction proceeds paid by the mobile network operators (MNOs).

Broadcasters choosing to participate in the auction would have different options depending on their current RF channel assignment, which involve different business and strategic tradeoffs. The incentive auction may represent a financial opportunity for broadcasters who remain on the air through the channel sharing option. Doing so, broadcasters would continue their business through a channel sharing agreement and they would not lose must-carry rights over cable and satellite. Depending on the agreement, channel sharing may lower both operating and capital expenditures for the broadcasters involved. To participate in the incentive auction, broadcasters will be offered the possibility to:

- Relinquish their current RF channel and share an RF channel with another broadcaster.
- Move from the UHF to the VHF band.
- Move from high VHF to low VHF.
- Go off the air.

The process combines two separate but linked auctions: a *reverse auction*, which will identify the prices at which broadcasters are willing to relinquish their RF channels; and a *forward auction*, which will determine the price MNOs are willing to pay to acquire the new frequencies. The two auctions will determine the buyers and sellers and also the amount of spectrum to be cleared in the 600 MHz band after reorganizing the television (TV) stations that remain on air. This process is known as *repacking* and will create contiguous blocks of cleared spectrum at the high frequency side of the UHF band for mobile use.

This article presents an overview of the upcoming TV broadcast spectrum incentive auction in the U.S. and reviews the main business, regulatory, and technical challenges of a successful incentive auction. The article presents the potential plans for the 600 MHz band, which could repurpose up to 144 MHz of broadcast TV spectrum and allocate 120 MHz for mobile broadband. The article discusses the broadcast frequency repacking and the opportunities that may bring a migration to the new and improved global DTT standard that is currently in development, known as ATSC 3.0.

Incentive Auctions Overview

Incentive auctions were introduced in the 2010 U.S. National Broadband Plan [3]. After a few delays, the auction may take place in early 2016, although it should be pointed out that at the time this article was written the FCC was working to determine the final auction rules and procedures. There are several outstanding issues left to be resolved, such as the determination of the initial broadcast spectrum clearing target, opening bid prices, benchmarks for the final stage rule, and the final TV channel assignment process.

Incentive auctions are a market-driven tool for repurposing spectrum. The idea behind the incentive auction is that broadcasters may be willing to voluntarily relinquish all or some of their spectrum usage rights during the reverse auction in exchange for a share of the income raised in the forward auction. Broadcasters may, of course, elect not to participate in the incentive auction. Current estimates, while varying widely, indicate that about 10.7 million TV households in the U.S. rely on DTT only, which represents an approximately 10 percent penetration, although the percentage of households with at least one DTT set is considerably larger.

In order for the incentive auction to be carried through to completion, it is required to raise substantial proceeds. Different estimates based on previous spectrum auctions predict that the proceeds could approach up to USD $45 billion [3]. Compensation may need to be significant for broadcasters not only in the top markets, but also in mid to smaller-sized markets in order to clear spectrum on a nationwide basis.

The incentive auction is structured in two separate but interdependent auctions, known as reverse and forward auctions, linked by the repacking process. For a given spectrum clear target defined by the FCC:

- The **reverse auction** will determine the price at which broadcasters will voluntarily relinquish their spectrum usage rights and the amount of spectrum available in each market.

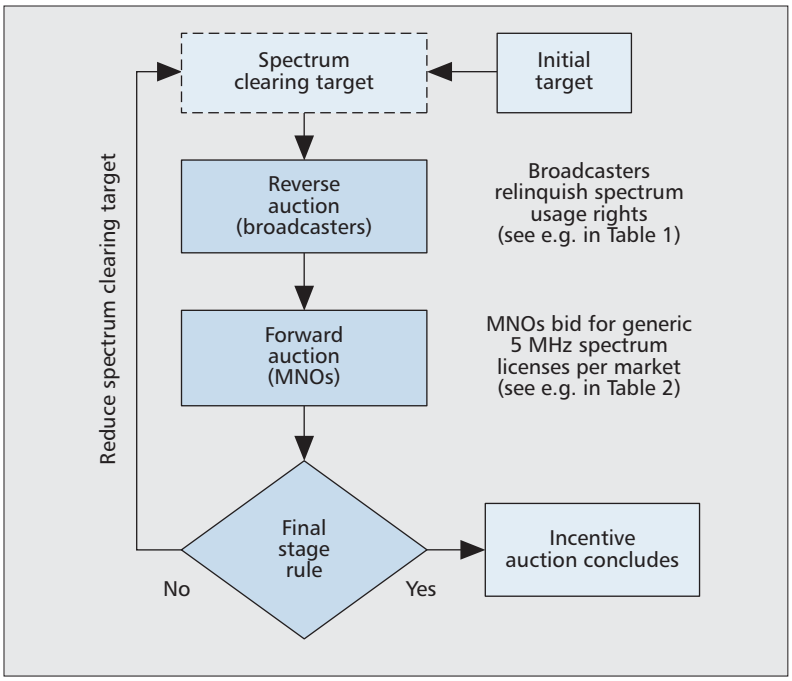


Figure 2. Integration of the reverse and forward auctions.

- The **forward auction** will determine the price MNOs are willing to pay for such spectrum.

The **repacking process** consists of reorganizing and assigning RF channels to the broadcast TV stations that remain on air after the incentive auction in order to create contiguous blocks of cleared spectrum at the upper frequency range of the 600 MHz band for mobile broadband use.

The participation of the broadcasters in the incentive auction is totally flexible and voluntary, and broadcasters that participate also have the possibility to drop out of the auction (these broadcasters will be treated as if they had not participated in the auction). The FCC has devised a flexible scheme to encourage the participation of the broadcasters with four different options, which involve different business and strategic trade-offs, depending on their current RF channel assignment. The four options are:

1 Bid to relinquish a UHF channel to move to a high VHF (174 to 216 MHz) or to a low VHF (54 to 88 MHz) channel.
2 Bid to relinquish a high VHF channel to move to a low VHF channel.
3 Bid to relinquish their current UHF channel and share a channel with another broadcaster after the auction.
4 Bid to relinquish their license and go off the air.

Only broadcasters who relinquish their spectrum usage rights will share the auction proceeds based on the value of their spectrum. Broadcasters not relinquishing their spectrum licenses will continue operating in the same frequency band, be subject to a potentially substantial disruptive RF channel reassignment resulting from the repacking process, and be entitled to reimbursement for relocation costs from a USD $1.75 billion TV broadcaster relocation fund, which must be generated from the proceeds of the forward auction. It is intended that funds would cover up to 80 percent of the eligible costs for commercial operators and up to 90 percent for non-commercial stations. Broadcasters that continue over-the-air operation after the auction will retain the mandatory carriage rights on cable and satellite systems.

Broadcasters sharing an RF channel must retain the capability to transmit at least one standard definition (SD) TV channel. This is the only constraint imposed by the FCC, and the actual terms of the sharing agreement are left to the broadcasters. Channel sharing agreements should be executed prior to the auction.

Regarding the possibility of migrating from the UHF down to the VHF bands, broadcasters will be able to choose whether to bid for the high-VHF or the low-VHF band, but the actual RF channel will be assigned in the repacking process if the broadcaster's bid is selected in the auction.

Broadcasters relinquishing their license or accepting a channel sharing bid would have to vacate their RF channels within three months after receiving the economic compensation. Broadcasters remaining on the air would have up to 39 months after the auction ends to move to the new RF channel assigned in the repacking process.

The reverse and forward auctions will be integrated in a series of stages with a reserve price mechanism such that market forces will determine the amount of cleared spectrum and the revenues raised. Each stage will consist of a reverse auction followed by a forward auction for a given spectrum clear target defined by the FCC. An extended round may be performed within the forward auction if the final stage rule is not met but bidding stops in high-demand markets. Figure 2 illustrates the integration of the two auctions.

The opening prices are the starting prices for the reverse auction, and they are the highest prices offered to TV stations for the three options to relinquish UHF spectrum usage rights: go off the air, move to a low VHF channel, or move to a high VHF channel. The opening prices will be published by the FCC before the auction, and it is expected that they will be very high in order to gather interest from broadcasters. Prices will be calculated systematically for each station, taking into account its covered population and the potential interference in the repacking process, and a uniform price (base clock price) for a UHF station going off the air. For moving to low VHF and high VHF, tentative price ranges are between 67 and 80 percent, and between 33 and 50 percent of the station's price to go off-air, respectively.

The initial spectrum clearing target for the first round will be determined based on the initial position of the broadcasters indicated in the pre-auction application process. The reverse auction will start at the highest clearing target possible based on the eleven potential scenarios for the 600 MHz band plan described later.

The first reverse auction bidding process will determine the total amount of incentive payments to broadcasters required to clear the initial clearing target. If the proceeds of the forward auction cover the bids of the reverse

[1] MHz-POP is calculated as follows: the total price paid for a license is divided by the product of the bandwidth in MHz times the population covered by the license; the FCC initially proposed an average price per MHz-POP benchmark of $1.25 in the largest 40 markets by population [3].

auction, the TV broadcaster relocation fund (USD $1.75 billion) [3], the auction costs of the FCC, and the average price per MHz-POP[1] satisfies the minimum requirement established by the FCC before the auction, then the incentive auction will close at the initial spectrum clearing target. If not, additional rounds will be run with progressively lower spectrum targets in the reverse auction, and consequently less spectrum available in the forward auction, until the final stage rule is satisfied. An extended round may be performed within the forward auction if the final stage rule is not met but bidding stops in high-demand markets in order to give MNOs the opportunity to express demand at higher prices to meet the final stage rule, avoiding the need to move to a lower clearing target.

The Reverse Auction

The reverse auction will follow a descending clock auction format, in which prices are progressively reduced until the number of RF channels that broadcasters are willing to move or relinquish matches the clearing target. In each bidding round, broadcasters will be offered prices for one or more bid options, and they will have to indicate their choices at those prices. The descending clock structure allows broadcasters to react to the prices provided by the reverse auction, rather than having to formulate their own bids.

Table 1 illustrates an example of the reverse auction with a clearing target of two stations in a given market. The opening price is $$$$, and five broadcasters (stations) accept this price. For the sake of simplicity, all stations are assumed technically identical. Since there are more stations than required (i.e. excess of demand), a new round of bidding is carried out with a lower bid price (in the example, $$$). Only three broadcasters accept this price, but since there is one station above the clearing target, another round is performed with a lower price of $$. Only one broadcaster accepts this price, and hence the two broadcasters that accepted a price of $$$ will have to indicate the price level between $$ and $$$ (i.e. between the opening and closing prices of a round) that they accept to relinquish their station. This methodology is known as intra-round bidding, and it will allow broadcasters to drop out of the auction and remain on air if their target price is not satisfied.

From the example, it can be noted that not all stations willing to accept an offered price will be selected to relinquish their licenses. Furthermore, stations in practice are not technically identical, as assumed in the example. This means that some stations are more suitable than others to fulfill a given spectrum clearing target, which also depends on the actual stations that participate in the auction. Hence, the reverse auction system will take into account the feasibility of repacking the remaining TV stations after the auction when offering the prices to each station. If a station cannot be assigned an RF channel, its price will become frozen, and will not be adjusted downward in the subsequent rounds. Otherwise, stations will continue to be offered reduced prices as long as they become necessary to meet the spectrum clearing target.

Round	Bid price	Stations accepting bid price	Stations in excess of clearing target
I	$$$$		+3
II	$$$		+1
III	$$		–1

Table 1. Illustrative reverse auction example where two stations are needed.

Round	Price per spectrum Block	Demand MN01	Demand MN02	Excess demand
III	$$$			–1
II	$$			+1
I	$			+3

Table 2. Illustrative forward auction example where two spectrum blocks (licenses) are available.

The Forward Auction

The forward auction will come after the reverse auction with an ascending clock auction format, in which prices are progressively increased until the number of spectrum licenses demanded by the MNOs matches the available spectrum (i.e. prices start low and are adjusted upward). In short, prices will continue to rise until there is no excess demand for the available spectrum. Another important feature of the forward auction is that MNOs will bid on the desired number of generic paired 5 MHz spectrum blocks per market (i.e. a 5 MHz block for the uplink and another 5 MHz block for the downlink). After the auction, a separate auction round process will be carried to allocate the specific blocks per market to the MNOs.

Table 2 illustrates an example of the forward auction for three available spectrum blocks in a given market in which only two MNOs are interested to obtain new licenses. The opening bid price is $, and for this price the two MNOs demand three spectrum blocks each, hence having an excess demand of three blocks. Another round is carried out with a higher price per block ($$), which reduces the demand of both operators to two blocks. The third round increases the bid price to $$$, which reduces the demand of the two MNOs to one block each. The price for the third block would be the highest bid of the two MNOs between $$ and $$$ (i.e. the opening and closing prices of the round). That is, the same intra-round bidding scheme used in the reverse auction can be applied in the forward auction.

TV Broadcast Spectrum Repacking

Terrestrial broadcast TV in the U.S. employs 222 MHz of spectrum in the UHF band and 72 MHz in the VHF bands, with a total number of over 8000 TV stations. The frequency planning for the first-generation U.S. DTT standard ATSC was based on a multi-frequency network

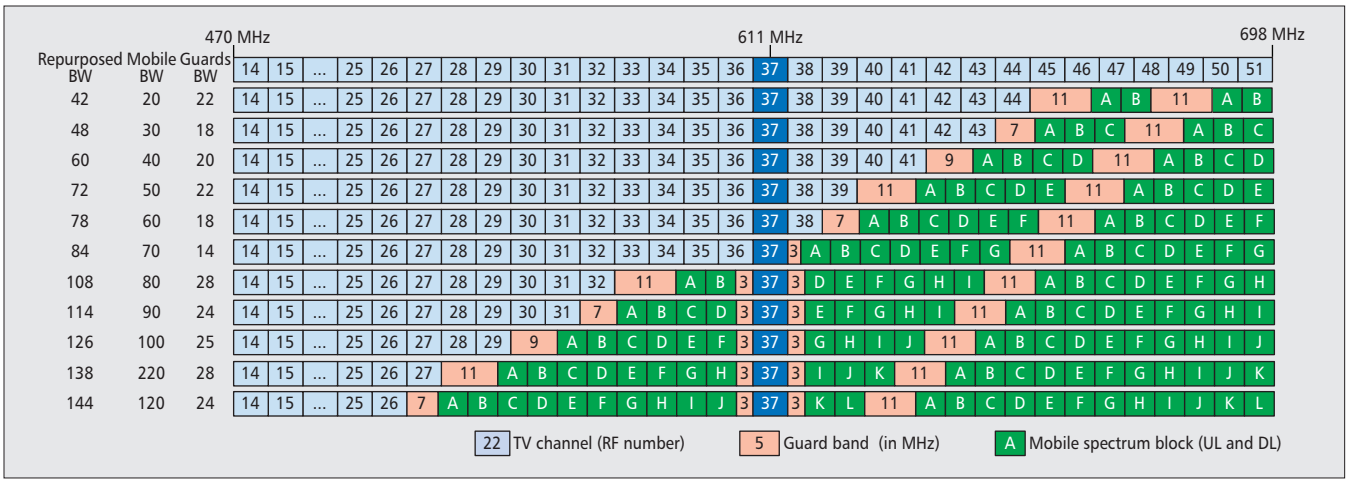


Figure 3. Potential scenarios for the 600 MHz band plan after the incentive auction.

topology, where each TV station is assigned a 6 MHz RF channel to cover a particular geographical area. In the U.S., TV stations are classified as full power ("*Full Power*"), class A low power television ("*Class A*"), low-power television ("*LPTV*"), or TV translator ("*Translator*"). Full-power stations are traditional high-power high-tower stations with up to 1 MW effective radiated power (ERP) in the UHF band, and have the highest priority regarding RF interference. LPTV stations are limited to up to 1 kW ERP in the UHF band, and also cover translators that retransmit the signals of a TV broadcast station. LPTV stations and translators are considered secondary to full-power stations, unless they are classified as class A. According to the FCC, there are 1782 full power stations, 465 class A low power stations, 1980 LPTV, and 4175 translators. It should be pointed out that only full-power (commercial and non-commercial) and class A licensees are eligible to participate in the incentive auction, and that licensees of LPTV and TV translator stations cannot participate.

As part of the reverse auction, the TV stations that remain on air would be reorganized so that they occupy a smaller portion of the UHF band toward the low frequency side allowing repurposing a contiguous portion of the high frequency side of the UHF band for mobile use. The terms of the repacking process have been designed to preserve the population and coverage served by each of the remaining individual TV stations as was determined for a baseline back in February 2012. A new RF channel assignment will not be allowed if the population served is reduced by more than 0.5 percent due to interference. The FCC is planning to use a software tool to determine repack feasibility. This tool makes use of two files. One file contains possible repack channels for each station that are based on restrictions associated with interference protection into Canada and Mexico and with protection of public safety allocations. The second file contains the channels for each station and for each of these channels the stations that would cause the co and adjacent interference over the 0.5 percent limit.

The 600 MHz Band Plan

The TV spectrum incentive auction requires a flexible band plan for the 600 MHz band because the quantity of broadcast spectrum that will be repurposed for mobile broadband will be an outcome of the auction itself. In this section we explain the flexible framework to account for different amounts of cleared spectrum. Ideally, after the reverse auction the same amount of spectrum will be available nationwide. The possibility of modifying the framework to accommodate varying amounts of available mobile spectrum per market is currently one of the open topics under discussion. Doing so would result in impaired spectrum blocks and a reduction in the number of available spectrum blocks in some markets. The FCC has initially proposed two categories of generic spectrum blocks: category 1 with at most 15 percent of the population impaired by interference, and category 2 with over 15 percent up to 50 percent of interfered population (spectrum blocks with more than 50 percent of impaired population will not be offered in the auction), having different price clocks for each category, and reducing the prices of the impaired blocks proportionally to the percentage of interfered population [3]. The final stage rule would only apply to category 1 blocks in the most populated markets, and hence after the final stage rule is satisfied the forward auction will continue until there is no excess demand in any category in any market.

Figure 3 shows the eleven potential scenarios that have been defined depending on the amount of spectrum cleared in the auction, ranging from a maximum of 144 MHz (24 RF channels) down to a minimum of 42 MHz (seven RF channels). The amount of spectrum reallocated to mobile services will range between 120 MHz in the most aggressive scenario and only 20 MHz in the most conservative scenario. The envisaged frequency plan for mobile services consists of a frequency division duplexing (FDD) scheme with specific paired uplink and downlink bands comprised of 5 MHz frequency blocks. The number of RF channels in the UHF band available for the repacking of broadcast TV licenses that remain on air would be reduced from 37 (from 14 up to

51, except RF channel 37, which is allocated to radio astronomy, RAS, and wireless medical telemetry services (WMTS)) down to 12 in the most aggressive scenario (30 in the most conservative scenario). The FCC has initially proposed 70 MHz as the forward auction spectrum benchmark, corresponding to a broadcast spectrum clearing target of 84 MHz [3]. It should be pointed out that the RF channel 37 will continue being reserved for RAS and WMTS applications after the incentive auction. In the figure it can also be noted that each scenario includes guard bands between the different technologies, including a duplex guard band of 11 MHz between the uplink and the downlink mobile transmissions. The guard band between broadcast and mobile services would range between 7 MHz up to 11 MHz, which is intended to avoid interference issues in adjacent channels [4].

In the U.S., unlicensed devices can operate on broadcast television RF channels that are not used at a given location, known as TV white spaces. Unlicensed devices are to use certified radio equipment, may not cause harmful interference to licensed incumbent services, and must accept any interference that they receive. Interference from TV white space devices into the incumbent services is supposed to be controlled through a spectrum database of protected incumbent service areas. With the new 600 MHz band, unlicensed devices and wireless microphones would be allowed to operate in TV white spaces in the repacked broadcast TV band, but also in the new guard bands of the resulting 600 MHz band plan. Furthermore, the FCC has proposed to allow unlicensed devices to operate for the first time on RF channel 37 by establishing interference protection margins for existing operations in the white space database. This proposal seeks to promote greater innovation in new products and services by enabling spectrum for unlicensed devices on a nationwide basis because currently the amount of available spectrum is very limited or even nonexistent in many major markets.

Challenges of the TV Spectrum Incentive Auction

The TV spectrum incentive auction is a very challenging process because all the pieces have to work together in order to be successful. For example, the opening bids by the FCC could break the auction from the beginning, as they will directly influence the participation of the broadcasters and the levels at which they might potentially drop out of the process. Efficient repacking of the remaining TV stations that remain on air after the auction must also be accomplished as a component of a successful auction. The repacking algorithm may be the most complex technical element of the overall process, which will be used in the reverse auction to help determine if a bid of a TV station is critical to the relevant spectrum clearing target.

Migrating from the UHF band down to the VHF band (and from the high-VHF down the low-VHF) may imply reception issues for the viewers. Although signals would reach further, building penetration loss would increase and, in order for them to maintain their efficiency, the size of customer receive antennas would need to increase.

The incentive auction will greatly affect the use of wireless microphones, low power TV stations (not classified as class A), and unlicensed TV white space devices, since all their operation would have to be modified after the repacking process. A main drawback of the incentive auction process is that broadcasters that desire to continue their businesses well into the future may not have enough spectrum available for introducing new technologies.

ATSC 3.0

From an alternate perspective, the incentive auction represents a tremendous opportunity for broadcasters who are interested in thriving in the business well into the future by embracing the technological advancements that are available and would greatly enhance the consumer experience. In comparison to the high costs of the repacking process ($2 million to $5 million per station), which may not be fully covered by the broadcast relocation fund, the additional costs to migrate to a next-generation system may not be significant ($250 thousand to $500 thousand per station).

The Advanced Television Systems Committee (ATSC) is currently developing its next-generation TV broadcasting system, known as "ATSC 3.0," which aims to become the reference DTT technology worldwide, leveraging all the progress beyond prior state-of-the art. The ATSC 3.0 system is being designed to include, among other features, higher system capacity to deliver a combination of emerging ultra-high definition (UHD), high frame rate (HFR), and high dynamic range (HDR) services and robust indoor and mobile reception. A primary goal is to simultaneously reach both fixed and portable devices. It is expected that ATSC 3.0 will allow transmitting more than 30 Mb/s in a 6 MHz RF channel for the same operation point of ATSC 1.0 (15 dB signal-to-noise ratio at 19.4 Mb/s) [5]. Therefore, the combination of ATSC 3.0 with the new video codec HEVC (high-efficiency video codec), which, theoretically, provides up to four times the compression gain with respect to MPEG-2 video coding (two times with respect to MPEG-4/AVC), and its one-to-everyone broadcast architecture, results in the most spectrally efficient mass media content delivery system.

One of the problems with the introduction of ATSC 3.0 may be the lack of spectrum to allow for a simulcast period with both the existing service and the new service, to allow users to progressively update their receivers either with a new TV set or a new set-top-box. One possibility may be for two broadcasters to share their two RF channels, such that one RF channel is used for ATSC 1.0 (since this configuration may allow for the delivery of two lower quality high-definition HD services), and the other for ATSC 3.0.

Regarding the channel sharing of the new ATSC 3.0 services, broadcasters may be initially more interested in delivering a reduced set of TV channels (e.g. one each assuming that two

From an alternate perspective, the incentive auction represents a tremendous opportunity for broadcasters who are interested in thriving in the business well into the future by embracing the technological advancements that are available and would greatly enhance the consumer experience.

The introduction of the ATSC 3.0 next generation digital terrestrial television system makes possible the delivery of higher quality services to an entirely new set of display devices and allows for the service to continue being a competitive media platform. Therefore, the migration to ATSC 3.0 should be taken into account as part of the overall incentive auction.

broadcasters pair up) with a very robust configuration (e.g. SNR in the order of 0 dB to 3 dB), aiming for robust indoor reception on TVs and tablets. Another possibility would be to deliver two higher quality HD fixed services per RF channel. In both cases, ATSC 3.0 would enable the transmission of new services once the simulcast period is concluded and more spectrum becomes available.

Outlook

The upcoming broadcast TV spectrum incentive auction in the U.S. will be a groundbreaking spectrum event worldwide. This innovative spectrum auction scheme is a market-driven tool for repurposing spectrum. With a successful incentive auction, the U.S. would be the first country to make the upper portion of the 600 MHz band available for mobile broadband.

The incentive auction is a very challenging process from both technical and economic points of view because in order to be successful the reverse and forward auctions and the repacking process must work seamlessly. The opportunities that the repacking component of the auction presents to TV broadcasters should not be disregarded. The introduction of the ATSC 3.0 next generation digital terrestrial television system makes possible the delivery of higher quality services to an entirely new set of display devices and allows for the service to continue being a competitive media platform. Therefore, the migration to ATSC 3.0 should be taken into account as part of the overall incentive auction.

References

[1] European Broadcasting Union (EBU), "Spectrum Fact Sheet," July 2014.

[2] P. Lamy, "Results of the Work of the High Level Group on the Future Use of the UHF Band (470–790 MHZ)," Report to the European Commission, Sept. 2014.

[3] Federal Communications Commission (FCC), "Learn Program about Incentive Auctions," FCC.gov/LEARN.

[4] M. Fuentes *et al.*, "Coexistence of Digital Terrestrial Television and Next Generation Cellular Networks in the 700 MHz Band," *IEEE Wireless Commun. Mag.*, vol. 21, no. 6, Dec. 2014.

[5] L. Michael and D. Gómez-Barquero, "Modulation and Coding for ATSC 3.0," *Proc. IEEE Broadband Multimedia Systems and Broadcasting*, Ghent, Belgium, 2015.

Biographies

David Gómez-Barquero (dagobar@iteam.upv.es) received a Ph.D. degree in telecommunications engineering from the Universitat Politècnica de València (UPV), Spain, in 2009. He is currently a senior researcher (Ramon & Cajal Fellow) at UPV's Institute of Telecommunications and Multimedia Applications (iTEAM), where he leads a research group working on the development of next-generation digital terrestrial broadcast technologies, and a guest research scholar at the New Jersey Institute of Technology (NJIT). He is the editor of the book *Next Generation Mobile Broadcasting* (CRC Press, 2013), and the vice-chairman of the Modulation and Coding Ad-Hoc Group of the standardization process of the next-generation U.S. TV broadcasting system ATSC 3.0.

M. Winston Caldwell, P.E. (winston.caldwell@fox.com) received his bachelor of engineering degree in electrical engineering from Vanderbilt University, and his master of science degree in electrical engineering from the University of Southern California. He is a licensed professional engineer in the state of California with more than 20 years of electrical engineering experience, specializing in RF propagation, wireless communications, and antenna design. He is currently Vice President, Spectrum Engineering and Advanced Engineering, for 21st Century Fox's Fox Networks Group, where he is involved with the exploration of new broadcasting opportunities, including development work on the ATSC 3.0 next-generation TV broadcast standard, acting as vice-chairman of the Waveform Ad-Hoc Group. He is an active participant in the ITU, IEEE, NAB, NABA, and SMPTE, where he has provided technical expertise in the determination of compatibility requirements between the established broadcasting services and unlicensed TV band, ultra-wideband, power-line transmission, and IMT-Advanced devices.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE

BIO-INSPIRED CYBER SECURITY FOR COMMUNICATIONS AND NETWORKING

BACKGROUND

Nature is Earth's most amazing invention machine for solving problems and adapting to significant environmental changes. Its ability to address complex, large-scale problems with robust, adaptable, and efficient solutions results from many years of selection, genetic drift and mutations. Thus, it is not surprising that inventors and researchers often look to natural systems for inspiration and methods for solving problems in human-created artificial environments. This has resulted in the development of evolutionary algorithms including genetic algorithms and swarm algorithms, and of classifier and pattern-detection algorithms, such as neural networks, for solving hard computational problems.

A natural evolutionary driver is to survive long enough to create a next-generation of descendants and ensure their survival. One factor in survival is an organism's ability to defend against attackers, both predators and parasites, and against rapid changes in environmental conditions. Analogously, networks and communications systems use cyber security to defend their assets against cyber criminals, hostile organizations, hackers, activists, and sudden changes in the network environment (e.g., DDoS attacks). Many of the defense methods used by natural organisms may be mapped to cyber space to implement effective cyber security. Some examples include immune systems, invader detection, friend vs. foe, camouflage, mimicry, evasion, etc. Many cyber security technologies and systems in common use today have their roots in bio-inspired methods, including anti-virus, intrusion detection, threat behavior analysis, attribution, honeypots, counterattack, and the like. As the threats evolve to evade current cyber security technologies, similarly the bio-inspired security and defense technologies evolve to counter the threat.

The goal of this feature topic is twofold: (1) to survey the current academic and industry research in bio-inspired cyber security for communications and networking, so that the ComSoc community can understand the current evolutionary state of cyber threats, defenses, and intelligence, and can plan for future transitions of the research into practical implementations; and (2) to survey current academic and industry system projects, prototypes, and deployed products and services (including threat intelligence services) that implement the next generation of bio-inspired methods. Please note that we recognize that in some cases, details may be limited or obscured for security reasons. Topics of interests include, but are not limited to:

- Bio-inspired anomaly & intrusion detection
- Adaptation algorithms for cyber security & networking
- Biometrics related to cyber security & networking
- Bio-inspired security and networking algorithms & technologies
- Biomimetics related to cyber security & networking
- Bio-inspired cyber threat intelligence methods and systems
- Moving-target techniques
- Network Artificial Immune Systems
- Adaptive and Evolvable Systems
- Neural networks, evolutionary algorithms, and genetic algorithms for cyber security & networking
- Prediction techniques for cyber security & networking
- Information hiding solutions (steganography, watermarking) and detection for network traffic
- Cooperative defense systems
- Bio-inspired algorithms for dependable networks

SUBMISSIONS

Articles should be tutorial in nature and written in a style comprehensible and accessible to readers outside the specialty of the article. Authors must follow the IEEE Communications Magazine's guidelines for preparation of the manuscript. Complete guidelines for prospective authors can be found athttp://www.comsoc.org/commag/paper-submission-guidelines.

It is important to note that the IEEE Communications Magazine strongly limits mathematical content, and the number of figures and tables. Paper length should not exceed 4,500 words. All articles to be considered for publication must be submitted through the IEEE Manuscript Central site (http://mc.manuscriptcentral.com/commag-ieee) by the deadline. Submit articles to the "June 2016 / Bio-inspired cyber security for communication and networking" category.

SCHEDULE FOR SUBMISSIONS

- Submission Deadline: November 1, 2015
- Notification Due Date: February 1, 2016
- Final Version Due Date: April 1, 2016
- Feature Topic Publication Date: June 2016

GUEST EDITORS

Wojciech Mazurczyk
Warsaw University of Technology
Poland
wmazurczyk@tele.pw.edu.pl

Sean Moore
Centripetal Networks
USA
smoorephd@gmail.com

Errin W. Fulp
Wake Forest University
USA
fulp@wfu.edu

Hiroshi Wada
Unitrends
Australia
hiroshi.wada@nicta.com.au

Kenji Leibnitz
National Institute of Information and Communications Technology
Japan
leibnitz@nict.go.j

5G Spectrum: Is China Ready?

Tan Wang, Gen Li, Jiaxin Ding, Qingyu Miao, Jingchun Li, and Ying Wang

ABSTRACT

With a considerable ratio of the world's mobile users, China has been actively promoting research on 5G, in which the spectrum issue is of great interest. New 5G characteristics put forward a lot of requirements for spectrum in terms of total amount, candidate bands, as well as new challenges for spectrum usage methods and management. Based on China's current situation, this article first discusses the 5G vision, spectrum demands, and potential candidate bands. Furthermore, it is indicated that spectrum sharing will bring many benefits for 5G systems, and different sharing scenarios are summarized. Finally, based on the current framework of spectrum management in China, potential services classification and spectrum assessment are proposed to accommodate new 5G requirements.

INTRODUCTION

1.286 billion was the total number of mobile subscribers in China at the end of 2014. The penetration ratio has exceeded 94.5 percent of the general population. For China, the country with the largest number of mobile users on the planet, the past year witnessed dramatically fast development of fourth generation (4G) industry. In October 2014, 3G subscribers saw negative growth for the first time. However, the number of mobile broadband subscribers (i.e., 3G plus 4G users) exceeded 582 million by the end of the year. Meanwhile, mobile Internet traffic maintains high-speed growth. Based on a 47 percent annual increase, the average user traffic has reached 205 MB per month, 86.8 percent of which is from mobile phones [1].

But this is more than enough. People's demand for mobile data services knows no limit. Historically in the development of mobile communication systems, it has taken about 10 years to give birth to a new generation. As 4G commercializes, the research and development of 5G has been launched, targeting commercial deployments in 2020 and beyond, according to International Telecommunication Union Radio Communication Standards Sector (ITU-R) Working Party 5D [2].

As the fundamental carrier of cellular mobile communications, radio spectrum resources have a decisive effect on the scale of industry development. Among 5G studies, the spectrum issue is one of the most important parts. How much spectrum does 5G need? Where can appropriate spectrum bands be found? How will they be efficiently used and managed? There are many open questions to be answered.

In this context, global 5G research institutions pay great attention to research on spectrum. To name but a few, China's IMT-2020 Promotion Group (IMT-2020 PG) set up a spectrum working group to deal with 5G spectrum demand and candidate bands. Furthermore, the EU FP7 METIS project has delivered a dedicated spectrum report [3]. In addition, in some countries the released national spectrum strategy put the spectrum for the next generation of mobile communication in a prominent position, such as the report to U.S. President Obama from PCAST [4] and Ofcom's Spectrum Management Strategy from the United Kingdom [5].

This article discusses the related considerations on 5G spectrum in China. An overview is given of the 5G vision. We focus on the spectrum demand, followed by the potential candidate bands for 5G. New spectrum usage methods are discussed. Based on China's current framework of spectrum management, two new elements — service classification and spectrum assessment — are proposed. The summary and outlook are given in the last section.

WHAT IS 5G?

What requirements does 5G impose? Which new service scenarios should 5G support? These questions are being actively discussed in China as well as globally. Different organizations have given different answers, as summarized in Table 1.

The METIS project in Europe is the first public research project aimed at 5G. In METIS's 5G vision, large numbers of both human-centric and machine-type users will be provided with a wide variety of services [6]. This vision brings challenges to future 5G systems, such as very high data rates, super dense crowds of users, and improved end-to-end performance. The general Key performance indicators (KPIs) are listed in Table 1 in terms of relative parameters compared to today. Furthermore, to provide more specific research topics, 12 concrete test cases

Tan Wang, Jiaxin Ding, and Jingchun Li are with the State Radio Monitoring Center of China.

Gen Li and Qingyu Miao are with Ericsson Research.

Ying Wang is with Beijing University of Posts and Telecommunications.

0163-6804/15/$25.00 © 2015 IEEE

KPI item	4G ITU requirement	METIS (Europe)		5G Forum (Korea)	Future Forum (China)
		General	TC1		
Peak data rate	1 Gb/s	10 ×~100 ×	5 Gb/s at 20% space	50 Gb/s	10 Gb/s
Cell edge user data rate	6 Mb/s	10 × ~ 100 ×	1 Gb/s at 95% space	1 G b/s	100 Mb/s
User plane latency	10 ms	5 × reduced	10 ms	1 ms	1 ms
Mobility	350 km/h	350 km/h	6 km/h	350 km/h	500+ km/h
Connection density	—	—	0.1 million/km^2	—	1 million/km^2
Traffic density	—	1000 ×	0.1 Gb/s/m^2	—	0.01 Gb/s/m^2

Table 1. Summary of 5G KPIs from different organizations.

(TCs) are defined [6], such as the virtual reality office (TC1), where respective KPIs are given in Table 1. Meanwhile, many organizations in different areas have also started research on 5G. For example, the 5G Forum in Korea proposes more aggressive data rate requirements, as shown in Table 1.

In China, the Future Forum and IMT-2020 PG, as two main fora, are beginning to draw an overall picture for 5G based on China's situation [7, 8]. First, they envision that mobile Internet and the Internet of Things (IoT) are two main drivers of future mobile networks in China which will touch many aspects of life in the future (i.e., home, work, leisure, and transportation). Regarding the first three aspects, there is a common view on a global level in both Europe and China; thus, similar requirements are introduced in terms of data rate and latency.

However, the transportation aspect is different in China compared to other countries; the train is the most important transportation vehicle in China rather than the airplane as in western countries. A high-speed railway is becoming more and more popular throughout the country. Its maximum operation speed is already 350 km/h, and the experimental speed is 605 km/h, which implies that 500 km/h is very promising for 2020 when 5G will be deployed. Therefore, China regards transportation scenarios as important 5G use cases, and the mobility requirement is above 500 km/h. Considering China's large population, ultra-high traffic volume density and ultra-high connection density are real challenges for 5G systems, which are interpreted as absolute KPIs given in Table 1. In addition, spectrum efficiency and cost efficiency are envisioned to be enhanced by a factor of 100 compared to today's network.

Although they have different views on 5G, several organizations have a common understanding that 5G systems aim to meet totally different requirements in various extreme scenarios. Moreover, it is difficult to satisfy all requirements by one radio access technology (RAT). Therefore, 5G is supposed to be a bundle of new and heterogeneous technologies, such as the maturing of Long Term Evolution (LTE) and WiFi, as well as new RATs. As illustrated in Fig. 1, the overall 5G radio access solution will most likely consist of multiple well integrated RATs, where carrier frequencies range from low to extremely high frequencies. In addition, implementation, deployment, and compatibility issues change with carrier frequency due to different propagation characteristics.

How Much Spectrum Does 5G Need?

Every few years, ITU-R sets up agenda items in advance to study the future spectrum demand for international mobile telecommunication (IMT), and to support the consideration of additional spectrum allocations. Currently, ITU-R has almost finished study on IMT spectrum demand toward 2020. China is actively involved in these studies. Besides, China has also started study on spectrum demand beyond 2020.

Demand in 2020 China

How should IMT spectrum demands be calculated? Generally, a methodology starts with an analysis of future market and traffic volume, moves on to calculate and distribute the traffic on different RATs, and then calculates the required capacity before concluding the estimation. The actual calculation process can be very complicated when there are a variety of traffic types, different environments, and multiple cell types of different RATs. For example, imagine estimating the data rates of a high-quality video streaming user located in indoor offices, connecting with future 5G small cells, in 2025.

Many countries have made contributions to the calculation methodology. Some of them, such as China and the United Kingdom, focused on the improvements of the existing ITU-R method specified in Recommendation M.1768. There are also original methodologies proposed by the GSM Association (GSMA), U.S. Federal

Every few years, ITU-R sets up agenda items in advance to study the future spectrum demand for international mobile telecommunication (IMT), and to support the consideration of additional spectrum allocations. Currently, ITU-R has almost finished study on IMT spectrum demand toward 2020. China is actively involved in these studies.

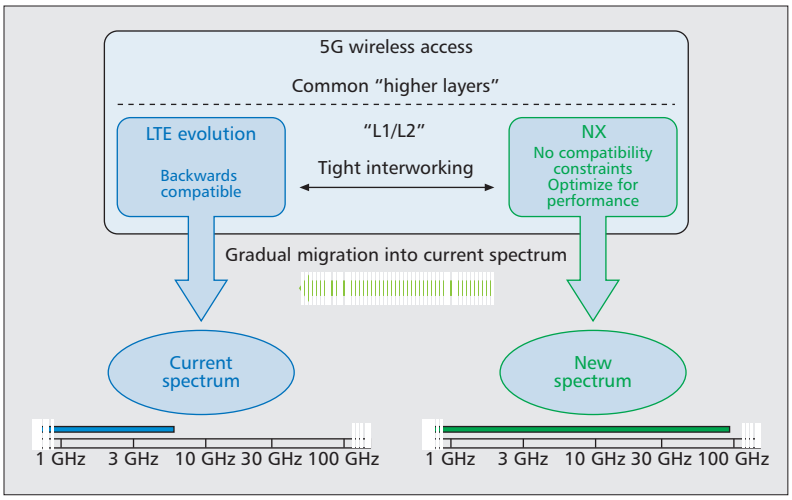


Figure 1. 5G wireless access in relation to spectrum. (Source: Ericsson white paper on 5G radio access.)

Communications Commission (FCC), Russia, and so on [9]. In China, besides using existing methodologies, a new method is also proposed based on national considerations.

The main point of the new method is to estimate the upper bound of demand. In fact, it is proved by the operational data that the area which has the largest spectrum demand is always in urban areas. When requirements of such a scenario are satisfied, the solution might be applicable for others as well. In the calculation, a typical hotspot zone in Beijing is selected as the research area. To gather the historical data within, the government issued an investigation letter to all network operators. The data survey is based on the operator's network management system in China, in which there is a major difference between the proposed method and other international methods. Supported by these first-hand data, analyses are made to estimate the traffic increase, traffic distribution, and base station (BS) deployments in the future. The general flow chart of the methodology is shown in Fig. 2. According to the calculation, the total IMT spectrum demand in China is 1350–1810 MHz in 2020 [10].

This demand is huge. At the end of 2014, China has already planned 687 MHz spectrum for IMT. There is at least a 663 MHz deficit. Moreover, in the period of 5G commercialization after 2020, the total spectrum demand may continue to increase.

Demand Considerations in 2020–2030

The above result is, to some extent, a total amount of spectrum, and the range depends on the assumptions made in the estimation process. However, in view of 5G beyond 2020, it is probably difficult to measure the new demand only in a total number. For example, in 5G indoor high traffic scenarios, in order to achieve high peak data rates, the demand for frequency bandwidth may be up to several gigahertz. This can be solved by using higher frequencies and denser deployments. But these solutions may not be the best choice for outdoor wide area scenarios. Therefore, it will be useful to separately estimate how much spectrum is required for coverage, capacity, performance, and connections for each 5G scenario to perform a mapping onto different frequency bands, such as bands below 1 GHz, between 1–6 GHz and even above 6 GHz. It is likely that the results from different scenarios will vary a lot. Generally, the 5G spectrum demand estimation will be a comprehensive outcome, indicating spectrum solutions for different scenarios. It is expected that 5G requires a higher the total amount, wider with respect to individual bandwidths, greater in range, and more flexible in usage pattern.

Where Can Spectrum Bands for 5G Be Found?

Is China ready to provide sufficient frequency bands for 5G? Overall, the potential bands can be divided into two parts: bands below and above 6 GHz [11].

Spectrum below 6 GHz

For wireless communications, lower frequencies provide better coverage. Currently, almost all countries use spectrum below 6 GHz for IMT systems. Besides achieving high data rates, it is also necessary to guarantee wide-area coverage and outdoor-to-indoor coverage in 5G. Therefore, spectrums below 6 GHz form a very important part of the 5G spectrum solution. In China, potential 5G spectra below 6 GHz include the following aspects.

Spectrum refarming: Until the end of 2014 in China, 2G, 3G, and 4G networks respectively occupy 132 MHz, 95 MHz, and 250 MHz spectrum total, respectively . For the two operators authorized for LTE hybrid network trial, China Telecom has 2 × 15 MHz spectrum for LTE FDD, while China Unicom only has 2 × 10 MHz in the beginning. Market competition has prompted Chinese Unicom to accelerate the pace of 2G spectrum refarming for 3G and 4G networks. On the other hand, being the world's largest TD-LTE network, China Mobile has also expressed willingness to refarm 2G spectrum for LTE frequency-division duplexing (FDD). When 5G is put into use, it is expected that some spectrum from the old generation could be refarmed for it as well. However, refarming does not increase the total spectrum amount.

Already identified IMT spectrum: In the Regulations of Radio Frequency Division of China, there are several frequency bands identified for IMT, specified in footnote CHN28 [12], such as 2300–2400 MHz and 3400–3600 MHz. For 2300–2400 MHz in China, after careful compatibility study, IMT systems have proven the capability of coexisting with radio location services, but limited to only indoor use. This band was assigned in December 2013 for deploying TD-LTE systems.

Nevertheless, other bands are still to be studied with respect to compatibility before official use. For example, in China, 3400–3600 MHz is already used as extended C band for satellite services, since it can provide better propagation characteristics against the rain attenuation than higher frequency bands. Therefore, until now,

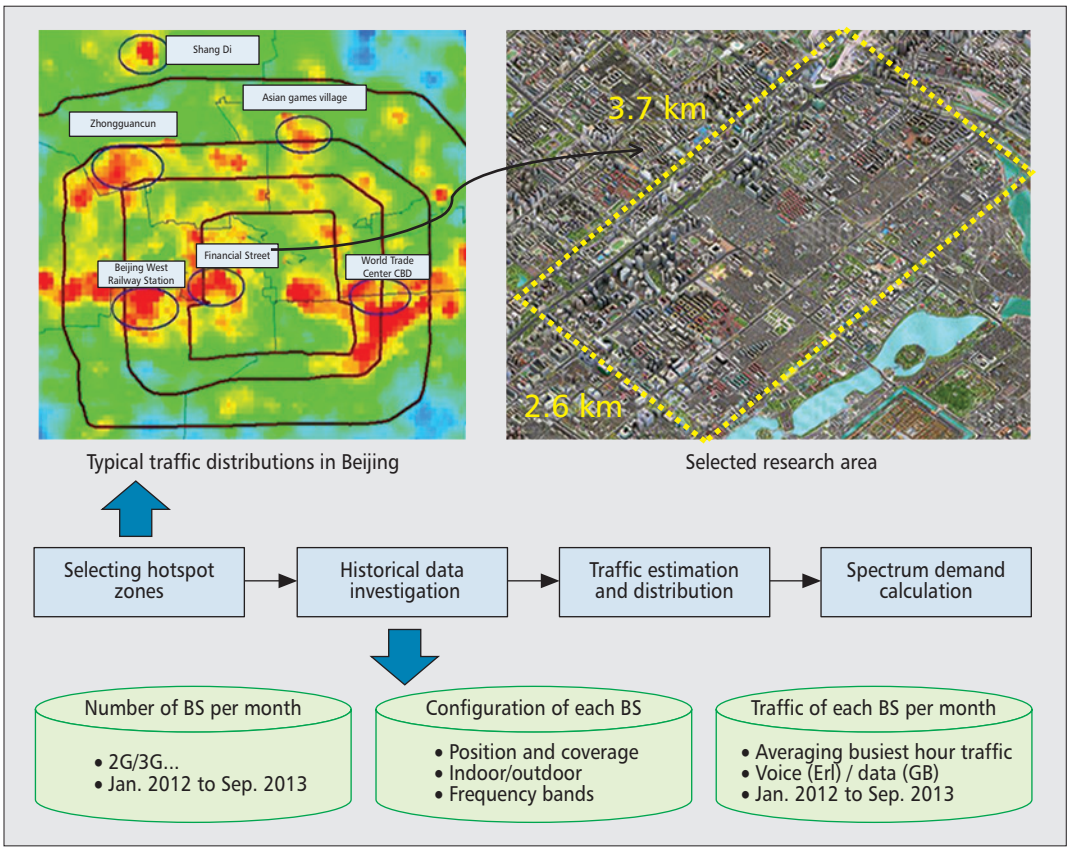


Figure 2. Flow chart of the spectrum demand methodology proposed by China, where a 3.7 km × 2.6 km square area around Beijing Financial Street is selected as the typical dense traffic area.

Spectra below 6 GHz are the best resources for IMT in the near future. However, due to their scarcity in China and the increasing difficulty in realizing international harmonization, it is time to seek spectrum above 6 GHz. Another way to utilize the benefits of these spectra is probably to share with other radio services.

allowing this band for IMT systems needs further demonstration and coordination.

Candidate bands being studied in ITU-R: In the current ITU-R research cycle, there are dozens of candidate frequency bands proposed for IMT from different nations [9]. They include, but are not limited, to 3300–3400 MHz, 4400–4500 MHz, and 4800–4990 MHz, which are currently supported by China. Other regions may have different considerations. For example, Europe supports 1427–1452 MHz, 1452–1492 MHz, 3400–3600 MHz, and 3600–3800 MHz; the United States supports 470–694 MHz and 1695–1700 MHz. Therefore, to realize global harmonization, it is necessary to achieve further regional convergence.

Spectra below 6 GHz are the best resources for IMT in the near future. However, due to their scarcity in China and the increasing difficulty in realizing international harmonization, it is time to seek spectrum above 6 GHz. Another way to utilize the benefits of these spectra is probably to share with other radio services, will discussed later.

Spectrum above 6 GHz

Above 6 GHz, there is a broad range of spectrum that could be considered for 5G. Here, it is easier to find a relatively "clean" band with much greater and continuous bandwidth (> 500 MHz). However, a lot of work is needed in seeking new bands for 5G above 6 GHz. Preliminary studies show that a lot of potential suitable frequency ranges could be found from 6 GHz to 100 GHz in China [11]. From the regulatory perspective, it is vital to perform solid research on these bands, including channel measurements, system modeling, and detailed studies on compatibility with currently used services [13].

How Can the Spectrum in 5G Be Used?

As we know, current 2G, 3G, and 4G systems in China only use licensed dedicated spectrum. However, according to the above analysis, 5G systems will need significantly more spectrum than today, and it is hard to find enough bands. Thus, using only licensed mode may not meet 5G requirements very well. To this end, this article provides a complete summary of spectrum sharing scenarios as the following categories:

1. *Vertical sharing* refers to spectrum sharing between users of different priority (e.g., primary and secondary); that is, users have unequal rights to spectrum access.
2. *Horizontal sharing* is sharing between systems that have the same priorities with respect to spectrum (i.e., different users have fair access rights to the spectrum). If users sharing the spectrum adopt the same technology, it is called *homogeneous horizontal sharing*; otherwise, it is referred to as *heterogeneous horizontal sharing*.

First, as shown in Fig. 3, the above-mentioned spectrum sharing scenarios are mapped to different types of spectrum: licensed, licensed shared access (LSA), and unlicensed.

In LSA mode, a licensee has the right to access spectrum that is unused by an incumbent user at certain locations and/or times. This vertical sharing is based on well defined conditions that are parts of a sharing license.

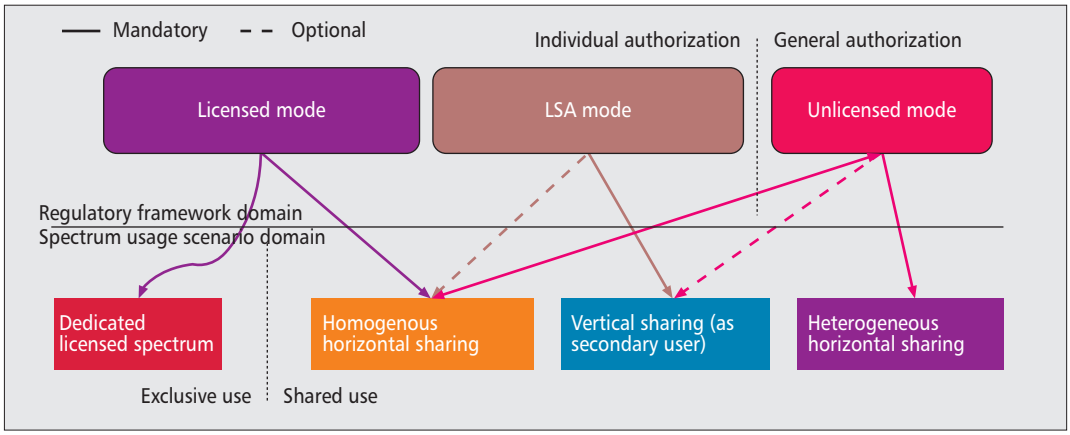


Figure 3. Illustration of spectrum type and related usage scenarios.

Licensed Spectrum

In licensed mode, spectrum is allotted to wireless systems for primary use. The only relevant spectrum sharing scenario is homogeneous horizontal sharing (i.e., sharing spectrum with other operators using the same RAT), which is also called co-primary or inter-operator spectrum sharing.

LSA Spectrum

In LSA mode, a licensee has the right to access spectrum that is unused by an incumbent user at certain locations and/or times. This vertical sharing is based on well defined conditions that are parts of a sharing license. In initial regulatory frameworks (e.g., the LSA concept developed in CEPT working group FM53), the licenses are expected to be long-term and exclusive. As it evolves, the LSA concept may in the longer term be combined with homogeneous horizontal sharing so that the issued licenses could be non-exclusive and require several licensees to coexist.

Unlicensed Spectrum

In unlicensed mode, a wireless system has to share spectrum with other unlicensed systems. For heterogeneous horizontal sharing in an unlicensed band, a system must be prepared for coexistence with any other technology that may be present in the band. Meanwhile, homogeneous horizontal sharing between different systems with the same RAT is unavoidable. Vertical sharing functionality may also be required for a system operating in unlicensed mode if a primary user exists in the band.

5G systems have some natural characteristics for facilitating the above-mentioned spectrum sharing techniques, which are explained in detail as follows.

•*Vertical sharing* can be a good way to unlock many bands for 5G systems as soon as possible. As mentioned above, the candidate bands for 5G systems are not "clean" even for frequencies above 6 GHz. Almost all candidate bands have other existing allocations and are already in use (e.g., fixed services or radar systems). It is not easy to coexist with them in every case. Thus, one way to use these bands for 5G systems is to repurpose them for licensed cellular usage. In some cases this approach may require a lot of effort and time to succeed. In order to make them available for 5G use in a relatively short time, one possible compromise is to introduce LSA. 5G systems would then operate under the constraint of protecting the incumbent service. To allow for efficient operation, a spectrum database or spectrum sensing technique may be employed. For any approach, quality of experience (QoE) for services of both primary and secondary systems should be guaranteed. An innovative public-private radio interference management framework that enables near-term spectrum sharing is proposed to guarantee 5G performance and user QoE [14].

•*Homogeneous horizontal sharing* is feasible for 5G systems particularly in high frequency band (i.e., above 6 GHz). Using these bands might not target providing ubiquitous coverage, but rather non-contiguous coverage islands. Furthermore, different operators may be required to serve different traffic volumes at the same instant. In such cases, spectrum sharing techniques can bring potential gains from statistical multiplexing. Besides, propagation conditions at high frequencies combined with high-gain beamforming for 5G systems is a promising approach for making good use of such frequencies.

•*Heterogeneous horizontal sharing* is likely to be employed only in unlicensed spectrum and can be used to boost capacity for 5G systems. There is currently and will also be a large amount of unlicensed spectrum even in high frequency bands (e.g., 59–64 GHz in China). The use of unlicensed spectrum for 4G systems is being discussed in the Third Generation Partnership Project (3GPP) in combination with licensed spectrum for critical control signaling. It is very likely that unlicensed mode will continue to be a complementary method of using spectrum for 5G systems.

Finally, it should be clear that applying these spectrum sharing techniques in practice needs additional functions. For example, vertical sharing may need databases to provide information about primary systems, while horizontal sharing may require coordination among different operators. These approaches may involve signaling overhead and implementation limits, which means extra effort is required. Thus, in our view, *licensed dedicated spectrum will continue to be the dominant spectrum usage method for 5G systems*

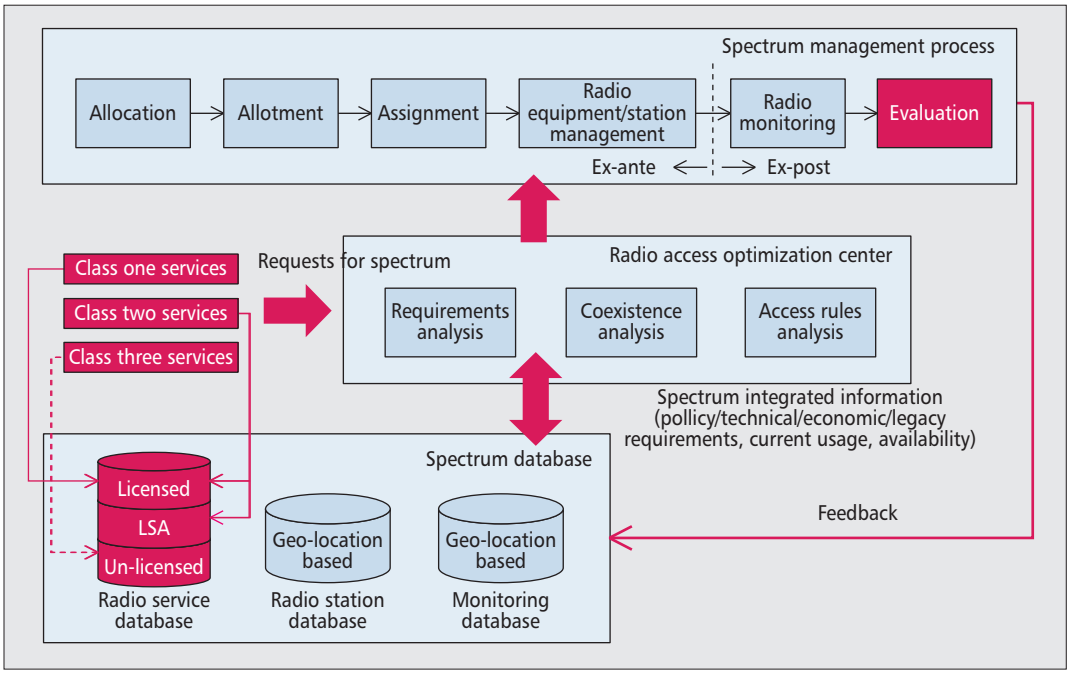


Figure 4. Proposed new elements in the current spectrum management framework in China.

When a new service access request arrives, the center checks the database, performs integrated analysis, and outputs the solution to support the process, which conversely inputs the information to the database for further use, establishing a closed-loop management.

due to better means of controlling interference and guaranteeing coverage, while other spectrum sharing approaches will act as complementary spectrum usage method when it is proved that there is benefit in future research work.

How to Manage the Spectrum in 5G?

At present, the spectrum license pattern in China mainly includes both licensed and unlicensed use. In addition, there is a light license mode, in which the user only needs to report but not apply for setting up a new radio station to the government for recording.

In such a management framework, spectrum sharing is not well supported. As 5G brings many more requirements in spectrum demand and also new spectrum usage methods to accommodate the requirements, it seems that spectrum management, to some extent, has to keep pace with innovation. In this section, based on the current framework of spectrum management in China shown in Fig. 4, two new elements are proposed: services classification and spectrum assessment. *To be clear, the new elements proposed in this article do not represent the current policy in China and are considered as research work aiming to give some food for thought. But from various aspects of the related work described below, they could be a possible trend.*

The *basic spectrum management* process is shown at the top of Fig. 4. The spectrum allocation (to radio services), allotment (to users or systems), and assignment (to specific radio stations), as well as the radio equipment and station management belong to ex-ante management. The radio monitoring belongs to ex-post management, including radio occupancy measurements, signal parameter and transmitter inspection, illegal transmitter detection and finding, and so on. The *radio access optimization center*, composed of relevant organizations and institutions, is in charge of spectrum demand calculation, coexistence, and access analysis. The *spectrum database* stores all the information of radio services, stations, and monitoring reports, which indicate the current spectrum usage and availability. When a new service access request arrives, the *center* checks the *database*, performs integrated analysis, and outputs the solution to support the *process*, which conversely inputs the information to the database for further use, establishing a closed-loop management.

Services Classification

Intuitively, not every radio service can be shared with IMT systems. Therefore, it is vital to make clear which types of service can be put forward for sharing. Here, it is proposed to divide them into three classes.

Class one services are those involving government affairs, national security, and public safety. They usually occupy specified licensed spectrum for free and need strict protection. Their spectrum cannot be shared with others. Some special radio services, such as radio astronomy, may also belong to this class. In future 5G standards, if there are related service aspects, will probably belong to class one services.

Class two services include commercial radio services, and also general public and dedicated services. The spectra for public mobile communication in 5G are included. Licensed or LSA mode could be used here. The government would flexibly charge the class two services, depending on their degrees of commercialization and behaviors of radio occupancy. At present, the Chinese government already charges the expense of frequency occupation, with clarified costs for different types of radio stations. Mean-

5G brings a lot of new features for spectrum, and the management needs to facilitate them. It is a trend to make spectrum access more open while strengthening the ex-post management. However, there are still many related political, technical, economic, and legacy issues to be studied.

while, there are detailed free lists and preferential lists for dedicated use. According to the latest Draft Amendments to China's Radio Management Regulations [15], auctioning is introduced for spectrum allotment. Recently, the Ministry of Industry and Information Technology (MIIT) of China is officially promoting the study on the marketing management mechanism of spectrum. In the study, it is necessary to distinguish the spectrum that might be participating in the auction from others.

Class three services are dedicated for public free use. The access process might also need spectrum sensing, such as WiFi. However, the transmission power of this class is required to be low.

Spectrum Assessment

Spectrum assessment belongs to ex-post management. It is not only the basis on whether a frequency band can be allotted to a new service, but also a way of supervising its utilization efficiency. In [15], the assessment of the current spectrum usage is proposed in China for the first time. Based on assessment results, the government can adjust or even retrieve the spectrum allotment. Currently, the State Radio Monitoring Center and State Radio Spectrum Management Center are conducting studies on spectrum assessment methodology and also the development of related test instruments.

To make fair assessment, a scientific and effective KPI system must be established. The KPI mainly includes three aspects. One is radio monitoring related information, such as noise and radio occupancy measurement results in different scenarios and locations. Another aspect is radio stations related information, such as location, RF parameters, and related statistics. Furthermore, the KPI system also includes service related information, such as different quality of service (QoS) and protection requirements. It is important to study the feasibility of providing such information in a technology- and service-neutral way, and whether it allows for a practical and efficient spectrum evaluation process.

It might not be possible to come to a correct conclusion if a frequency band can be allotted to others in the absence of any necessary KPI. For example, for existing fixed satellite services (space to Earth), by monitoring the radio occupancy ratio on the ground, there might be no room for others. But after clarifying the information of satellite receivers from the database (e.g., the exact locations for a limited number of stations), there might be possibilities for 5G use in specified scenarios. Another example is IMT sharing with radio location services. By measuring the occupancy ratio of radars, it is difficult to know when and where they work. The analysis must be combined with the radar station locations and transmit power information to avoid blocking of IMT receivers.

Conclusion

Based on China's situation, different aspects of 5G spectrum issues are discussed. It is proposed that controlled spectrum sharing is an important way of reusing spectrum to complement current licensed dedicated spectrum, which is still the basis for operation of 5G systems. According to China's current spectrum management process, the article proposes two innovative improvements, which are services classification and spectrum assessment. 5G brings a lot of new features for spectrum, and the management needs to facilitate them. It is a trend to make spectrum access more open while strengthening the ex-post management. However, there are still many related political, technical, economic, and legacy issues to be studied.

Acknowledgments

This article is supported by Chinese National Key Project under Grant No. 2014ZX03001027 and No. 2015ZX03002008, National HighTechnology Research and Development Program ("863"Program) of China under Grant No. 2014AA01A707, and Beijing Natural Science Foundation (4132050).

References

[1] MIIT of China, "The Economic Operation of the Communication Industry in 2014," gov. rep., 2014; http://www.miit.gov.cn/n11293472/n11293832/n11294132/n12858447/16414615.html

[2] ITU-R Working Party 5D, "Meeting Report of Working Group General Aspects," Nov. 2014; http://www.itu.int/md/R12-WP5D-C-0836/en.

[3] ICT-317669 METIS project, "Intermediate Description of the Spectrum Needs and Usage Principles," Del. D5.1, Apr. 2013; https://www.metis2020.com/documents/deliverables.

[4] PCAST, "Realizing the Full Potential of Government-Held Spectrum to Spur Economic Growth," July 2012.

[5] Ofcom, "Spectrum Management Strategy — Ofcom's Strategic Direction and Priorities for Managing Spectrum over the Next 10 Years," Apr. 2014.

[6] ICT-317669 METIS project, "Scenarios, Requirements and KPIs for 5G Mobile and Wireless System," del. D1.1, Apr. 2013; https://www.metis2020.com/documents/deliverables.

[7] Future Forum white paper, "5G SIG white paper," Nov. 2014, http://www.future-forum.org/en/ac_list.asp?id=129.

[8] China IMT-2020 PG white paper, "5G vision and requirements," May, 2014; http://www.imt-2020.org.cn/en/documents/.

[9] ITU-R Joint Task Group 4-5-6-7, "Draft CPM Text for WRC-15 Agenda Item 1.1," Annex 3 to Document 4-5-6-7/715-E, Aug. 2014.

[10] H. Biao and W. Tan, "Spectrum Requirements Calculation Based on Urban Hotspot in China for Mobile Communication in the Year 2020," *ZTE Technology J.*, vol. 20,no. 2, 2014, pp. 5–10.

[11] Cooperative Report, "Future Generation Mobile Systems and above 6GHz Spectrum," State Radio Monitoring Center (State Radio Spectrum Management Center) and GSMA, Sept. 2014.

[12] MIIT of China, "Regulations of Radio Frequency Division of People's Republic of China," Feb. 2014, http://www.miit.gov.cn/n11293472/n11293832/n12843926/n13917072/15865839.html.

[13] Y. Wang, J. Xu, and L. Jiang, "Challenges of System-Level Simulations and Performance Evaluation for 5G Wireless Networks," *IEEE Access*, Dec. 2014, pp. 2:1553–1561.

[14] J. Mitola *et al.*, "Accelerating 5G QoE via Public-Private Spectrum Sharing," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 77–85.

[15] Legislative Affairs Office of the State Council P. R. China, "Soliciting Public Opinions for Draft Amendments to China's Radio Management Regulations," May 2014; http://www.srrc.org.cn/NewsShow9865.aspx.

Biographies

Tan Wang (wangtan@srrc.org.cn) is an engineer with the State Radio Monitoring Center of China. He participated in Chinese 4G spectrum planning, and candidate bands evaluation and testing, and has been involved in the research on IMT spectrum requirements estimation in 2020 China and 5G spectrum issues. He has published more than 30

papers and authored 13 national patents. He received a Ph.D. in communication and information systems from Beijing University of Posts and Telecommunications in 2012.

GEN LI (gen.li@ericsson.com) has been working as an experienced researcher at Ericsson Research since 2012. He is now involved in spectrum work as well as research and development of 5G systems. He has published more than 20 papers and authored 30 international patents. He received a Ph.D. in communication and information systems from Beijing University of Posts and Telecommunications in 2012.

JIAXIN DING (dingjiaxin@srrc.org.cn) works for the State Radio Monitoring Center of China as a senior engineer. He received a Ph.D. in in communication and information systems from Beijing University of Posts and Telecommunications in 2003 and a visiting scholar position in the Department of Electronic Engineering of Stanford University during April 2010 to April 2011. He has been the Chairman of the Aeronautical and Maritime Task Group of the Asia-Pacific Telecommunity Wreless Group (APT/AWG) since 2012.

QINGYU MIAO (qingyu.miao@ericsson.com) has been working as master researcher at Ericsson Research on the research and development of 3G, 4G, and 5G systems since 2001. He has authored several publications and more than 100 patents in the field of wireless communication. He holds both an M.Sc. and a Ph.D. from Beijing University of Posts and Telecommunications, China.

JINGCHUN LI (lijingchun@srrc.org.cn), as a professorate senior engineer, is the deputy director and chief engineer of the State Radio Monitoring Center of China. He has over 30 years of research and working experience in radio monitoring, digital signal processing, and new radio technologies. He is a part-time professor at Beijing University of Posts and Telecommunications and Dalian University of Technology. He holds a Ph.D. in electromagnetic theory and microwave engineering from Xidian University, China.

YING WANG (wangying@bupt.edu.cn) received her Ph.D. in circuits and systems from Beijing University of Posts and Telecommunications (BUPT) in 2003. Now she is a professor and director of Radio Resource Management (RRM) Lab, Wireless Technology Innovation Institute, BUPT. Her research interests are in the area of cooperative and cognitive systems, and RRM in 5G. She took part in the Chinese Evaluation Group as a representative of BUPT. She has published more than 100 international papers.

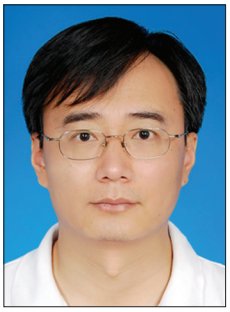
EMERGING APPLICATIONS, SERVICES, AND ENGINEERING FOR CELLULAR COGNITIVE SYSTEMS: PART II

Muhammad Zeeshan Shakir

Octavia A. Dobre

Muhammad Ali Imran

Apostolos Papathanassiou



Zhongshan Zhang

Athanasios V. Vasilakos

Honggang Wang

Hiroshi Harada

We are back this month with a sequel to our May 2015 Feature Topic on Emerging Applications, Services, and Engineering for Cognitive Cellular Systems (EASE for CCS). In this followup, we have eight more articles that present emerging applications integrated with CCS through novel frameworks. These emerging applications include energy harvesting, network virtualization, machine learning approaches, backward-compatible cognition in LTE, self-organization capabilities, integration of vehicular networking over unlicensed spectrum, 2.4/5 GHz hybrid architecture, and integration of D2D communications.

Cognition and self-organization in future networks are now widely considered as striking solutions to develop a sustainable cellular infrastructure. The first article, by Zaidi *et al.*, presents a novel framework to evaluate the performance of a cognitive metro-cellular network powered by harvesting a green energy resource, such as sunlight. The authors introduce useful formulations and a methodology, along with a case study, to show how we can assess the potential of self-sustained operation of cellular networks while keeping the "energy outage probability" at a minimum. The framework can be extended and exploited to study the potential of other green energy sources for the operation of cognitive metro-cellular networks.

The second article, by El-Sawy *et al.*, presents a vision of a more flexible and adaptive cognitive network architecture for 5G cellular networks. In this envisioned architecture, the network functionalities are virtually instantiated in the cloud while keeping in mind the application requiring the functionality, as well as the corresponding constraints and physical targets for those applications. The authors present a case study to highlight the potential benefits of their proposal. The framework proposed to evaluate the potential benefits of this architecture can be further employed in future literature and research on network function virtualization and software-defined networking-based architectures for the next generation of cognitive cellular networks.

The third article, by Gao *et al*,. introduces a multi-parameter cognitive architecture, consisting of parameter structuralization, cognition, and prediction. While previous studies ignore the potential relationships between different parameters, their structuralization can lead to cognition accuracy and reduced complexity. Parameter cognition can have sufficient or insufficient prior knowledge, and learning techniques are considered for the latter to intelligently establish a cognition knowledge base. Representative examples of how signal processing and machine learning techniques can be used to cognize multiple parameters are provided. Finally, following multiple parameters cognition, the internal structure and underly-

ing pattern of parameters can be used to predict their evolution. The proposed architecture represents a step forward to the ultimate goal of "full cognition" envisioned by Mitola, and thus toward future wireless communication systems.

Long Term Evolution cognitive radio (LTE-CR) over unlicensed spectrum, such as the industrial, scientific, and medical (ISM) band and television white space, has been studied in Third Generation Partnership Project (3GPP) standardization recently, in which network access is important in determining users' experiences. The fourth article, by Ling *et al.*, overviews network access schemes in existing CR networks first, and then proposes a system information block and cognitive pilot-channel-based backward-compatible scheme to facilitate the application of CR in LTE networks. This has potential to provide fast network access and put no stringent requirements on terminals. Moreover, due to the uncertainty of cognitive spectrum, which compromises the quality of service of users and the offloading efficiency of the network, the scheme is also load-aware. The article is expected to be of high reference value to the research and design of practical LTE-CR networks.

Small cells and their density in networks will be the trend of future wireless systems. Yuhua *et al.* investigate self-organizing optimization for cognitive small cells (CSC), which will play an important role in future cognitive cellular systems (CCS). The authors overview fundamental challenges and requirements for self-organizing optimization in CCS. A framework of game-theoretic solutions for self-organizing optimization in is established, and then featured game-theoretic models are proposed. A two-step scheme of game-theoretic solutions for self-organizing optimization is proposed. Some existing game models are introduced, and future research directions are presented. The authors discuss the usefulness of applying game theory for system analysis and optimization of cognitive small cell wireless networks.

The sixth article, by Shahid *et al.*, explores the possibility of using LTE device-to-device (D2D) communications for vehicular networks. Since a cellular network requires infrastructure-based communication, this incurs large delay that is unacceptable for vehicular applications. D2D offers an interesting solution to this problem by allowing vehicles to directly communicate without the need of an eNB. The article starts with a motivation to use LTE in vehicular networks followed by a review of D2D cellular communication and vehicle-to-vehicle (V2V) communication. Finally, the article proposes a cognitive algorithm to utilize LTE spectrum for V2V applications. Specifically, the algorithm uses path loss sensing to calculate the transmit power employed by vehicles, and allocates resources to vehicles according to the interference. Simulation results show a reduction in end-to-end delay and interference. The topic of the article is interesting and useful for future vehicular research. D2D communications could play a vital role in bringing vehicular communications into the mainstream.

The seventh article, by Chandra *et al.*, proposes the CogCell concept, which is a 2.4/5 GHz assisted 60 GHz picocellular network architecture, in which 60 GHz is used for high-speed data communication while 2.4/5 GHz WiFi is employed for control signaling. In CogCell, several 60 GHz picocells are managed by a single WiFi cell, thus facilitating easy and robust network and mobility management. In the absence of a 60 GHz connection, 2.4 GHz can be used as a fallback data communication option. By leveraging the sensing and processing capabilities of smart devices (e.g., motion sensors available in tablets and smartphones), cognitive and adaptive beam tracking can reduce the need for frequent re-beamforming at 60 GHz, and thus lead to more efficient spectrum utilization by effective switching between 2.4/5 GHz and 60 GHz bands for control and data communications. The proposed concept is supported by simulation results showing that with the help of rotation-vector sensor data, frequent re-beamforming for the 60 GHz directional links can be significantly reduced resulting in fewer requests to WiFi APs, thus demonstrating the efficient interplay between and 2.4/5 GHz WiFi and 60 GHz of the CogCell concept.

The eighth article, by Sakr *et al.*, highlights key challenges in resource allocation for in-band D2D enabled cellular networks and provides a comprehensive overview of the existing research advancements related to centralized and distributed resource allocation techniques. Since centralized solutions generally incur high computation and signaling overhead, distributed or semi-distributed solutions that exploit cognition at the D2D terminals are considered promising. Therefore, the authors propose a semi-distributed cognitive spectrum access (CSA) solution in which cognition at D2D terminals allows interference-aware decision making and limited control at the base stations to assist the D2D users in selecting the spectrum band with the least interference. The performance advantages of the proposed CSA scheme are analyzed quantitatively in terms of the channel access probability and spectral efficiency of cellular and D2D links. Finally, a number of directions for future research of CSA in D2D-enabled cellular networks are outlined.

In conclusion, EASE for CCS are now widely considered as frameworks to facilitate the heterogeneous demands of users in heterogeneous-type environments — particularly in the 5G network paradigm, where networks are anticipated to incorporate the provision of high-quality services to users with extremely low delays over limited spectral resources. The biggest challenge is to design new unified cross-layer network architectures for successful integration of the EASE frameworks by exploiting aggregation of highly distributed chunks of spectra (from unseen spectra to visible light spectra), network functions virtualization, spectrum harvesting, and orchestration of licensed and unlicensed spectral resources for ubiquitous connectivity.

Based on the high-quality contributions, we are sure that this two-part Feature Topic has been beneficial for further research, development, and technology advancement in future 5G CCS. Before closing this second part, we would like to thank all submitting authors for considering this Feature Topic as a potential venue for their research work,

reviewers for their high-quality evaluation, and the editorial/publishing team of *IEEE Communications Magazine* for their collaboration.

BIOGRAPHIES

MUHAMMAD ZEESHAN SHAKIR (muhammad.shakir@qatar.tamu.edu) has been an assistant research scientist at Texas A&M University at Qatar, Doha, since July 2012. He received his Ph.D. in electronic and electrical engineering from the University of Strathclyde, Glasgow, United Kingdom, in 2010. From January 2006 to September 2009, he was the joint recipient of an industrial research fund and a prestigious overseas research scholarship from the University of Strathclyde. His research interests include design and deployment of diverse wireless communication systems, including hyper-dense heterogeneous small cell networks. He has published more than 75 technical journal and conference papers, and has contributed to six books, all in reputable venues. He is co-author of two research monographs. Most of his research has been sponsored by Qatar National Research Fund and national industrial partners. He has served as a lead Guest Editor for *IEEE Communications Magazine* and *IEEE Wireless Communications*. He has served as co-chair of several special sessions/workshops and symposia at flagship conferences, such as IEEE ICC and GlobalSIP.

OCTAVIA A. DOBRE is an associate professor with Memorial University, Canada. In 2000 she was the recipient of a Royal Society scholarship in the United Kingdom, and in 2001 she held a Fulbright Fellowship in the United States. Her research interests include cognitive radio systems, spectrum sensing techniques, transceiver optimization algorithms, and dynamic spectrum access. She has published over 130 journal and conference papers in these areas. She is a Senior Editor for *IEEE Communications Letters*, and has served as Editor and Guest Editor for other prestigious IEEE journals. She has been the Co-Chair of technical symposia at flagship conferences, such as IEEE ICC and IEEE GLOBECOM.

MUHAMMAD ALI IMRAN is a reader (associate professor) in the Institute for Communication Systems, University of Surrey, United Kingdom. He is leading the physical layer work area for 5G innovation center and is curriculum design leader for the Engineering for Health center in Surrey. He has successfully led international projects encompassing the areas of energy efficiency, fundamental performance limits, sensor networks, and self-organizing cellular networks. He has supervised 20 successful Ph.D. graduates and published more than 150 peer-reviewed research papers. He is an Associate Editor for *IEEE Communications Letters* and the *IET Communications Journal*, and has served as a Guest Editor for other prestigious IEEE/IET journals.

APOSTOLOS (TOLIS) PAPATHANASSIOU is responsible for LTE PHY standardization and 5G technology development activities in the Next Generation and Standards (NGS) division of Intel's Communication and Devices Group (iCDG). He has more than 50 scientific contributions to international journals, conferences, and books, and more than 100 contributions to wireless standardization bodies such as 3GPP and IEEE 802.11/802.16. Previously at Intel, he led multiple standardization efforts in ITU-R and IEEE/WiMAX Forum. Before joining Intel, he worked on multiple-antenna PHY techniques and algorithms for 3G, WiFi, and satellite systems.

ZHONGSHAN ZHANG received his Ph.D. degree in electrical engineering in 2004 from Beijing University of Posts and Telecommunications. From February 2006 to March 2009, he was at the University of Alberta, Canada, as a postdoctoral fellow. He has also worked at DoCoMo Beijing Laboratories, Alcatel-Lucent Shanghai-Bell, and NEC China Laboratories as a researcher. He is currently a professor at the University of Science and Technology Beijing. His main research interests include self-organized networking, cognitive radio, and cooperative communications.

ATHANASIOS V. VASILAKOS is currently a professor with Kuwait University. He has served or is serving as an Editor for many technical journals, such as *IEEE Transactions on Network and Service Management*; *IEEE Transactions on Information Forensics and Security*; *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*; *IEEE Transactions on Information Technology in Biomedicine*; *IEEE Transactions on Computers*; *IEEE Transactions on Cloud Computing*; *ACM Transactions on Autonomous and Adaptive Systems*; and the *IEEE Journal on Selected Areas in Communications*. He is also General Chair of the Council of Computing of the European Alliances for Innovation (www.eai.eu).

HONGGANG WANG is an assistant professor at the University of Massachusetts Dartmouth. His research interests include wireless health, body area networks, cybersecurity, mobile multimedia and cloud, wireless networks and cyberphysical systems, and big data in mHealth. He serves as an Associate Editor of *IEEE Transactions on Big Data*, the *IEEE IoT* (Internet of Things) *Journal*, and the *IEEE Access Journal*, and an Associate Technical Editor of *IEEE Communications Magazine*.

HIROSHI HARADA joined the Communications Research Laboratory in 1995, part of Japan's Ministry of Posts and Communications (currently National Institute of Information and Communication Technology, NICT). Since 1995, he has researched software defined radio, cognitive radio, dynamic spectrum access network, and broadband wireless access systems on the microwave and millimeter-wave band. Currently he is director of the Smart Wireless Laboratory at NICT and has been a visiting professor of the University of Electro-Communications, Tokyo, Japan, since 2005.

Call for Papers
IEEE Communications Magazine

Wireless Communications, Networking, and Positioning with Unmanned Aerial Vehicles

Background

Enabled by the advances in computing, communication, and sensing as well as the miniaturization of devices, unmanned aerial vehicles (UAVs) such as balloons, quadcopters, and gliders, have been receiving significant attention in the research community. Indeed, UAVs have become an integral component in several critical applications such as border surveillance, disaster monitoring, traffic monitoring, remote sensing, and the transportation of goods, medicine, and first-aid. More recently, new possibilities for commercial applications and public service for UAVs have begun to emerge, with the potential to dramatically change the way in which we lead our daily lives. For instance, in 2013, Amazon announced a research and development initiative focused on its next-generation Prime Air delivery service. The goal of this service is to deliver packages into customers' hands in 30 minutes or less using small UAVs, each with a payload of several pounds. 2014 has been a pivotal year that has witnessed an unprecedented proliferation of personal drones, such as the Phantom and Inspire from DJI, AR Drone and Bebop Drone from Parrot, and IRIS Drone from 3D Robotics.

Among the many technical challenges accompanying the aforementioned applications, leveraging the use of UAVs for delivering broadband connectivity plays a central role in next generation communication systems. Facebook and Google announced in 2014 that they will use a network of drones which circle in the stratosphere over specific population centers to deliver broadband connectivity. Such solar-powered drones are capable of flying several years without refueling. UAVs have also been proposed as an effective solution for delivering broadband data rates in emergency situations through low-altitude platforms. For example, the ABSOLUTE, ANCHORS, and AVIGLE projects in Europe have been investigating the use of aerial base stations to establish opportunistic links and ad-hoc radio coverage during unexpected and temporary events. They can serve as a temporary, dynamic, and agile infrastructure for enabling broadband communications, and quickly localizing victims in case of disaster scenarios.

This proposed Feature Topic (FT) issue will gather articles from a wide range of perspectives in different industrial and research communities. The primary FT goals are to advance the understanding of the challenges faced in UAV communications, networking, and positioning over the next decade, and provide further awareness in the communications and networking communities on these challenges, thus fostering future research. Original research papers are to be solicited in topics including, but not limited to, the following themes on communications, networking, and positioning with UAVs.

- Existing and future communication architectures and technologies for small UAVs
- Delay-tolerant networking for cooperative UAV operations
- Design and evaluation of wireless UAV test beds, prototypes, and platforms
- Multi-hop and device-to-device communications with UAVs
- Interfaces and cross-platform communication for UAVs
- QoS mechanisms and performance evaluation for UAV networks
- Game-theoretic and control-theoretic mechanisms for UAV communications
- Use of civilian networks for small UAV communications
- Integrating 4G and 5G wireless technologies into UAV communications, such as millimeter wave communications, beamforming, moving networks, and machine type communications
- Use of UAVs for public safety and emergency communications, networking, and positioning
- Integration of software defined radio and cognitive radio techniques with UAVs
- Channel propagation measurements and modeling for UAV communication channels

Submissions

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions). Figures and tables should be limited to a combined total of six. The number of references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed if well-justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted athttp://www.comsoc.org/commag/paper-submission-guidelines. Please send a pdf (preferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "May 2016 / Wireless Communications, Networking and Positioning with UAVs" as the Feature Topic category for your submission.

Schedule for Submissions

- Submission Deadline: November 1, 2015
- Notification Due Date: January 15, 2016
- Final Version Due Date: March 1, 2016
- Feature Topic Publication Date: May 2016

Guest Editors

Ismail Guvenc
Florida International Univ., USA
iguvenc@fiu.edu

Walid Saad
Virgina Tech, USA
walids@vt.edu

Mehdi Bennis
Univ. of Oulu, Finland
bennis@ee.oulu.fi

Christian Wietfeld
TU Dortmund Univ., Germany
christian.wietfeld@tu-dortmund.de

Ming Ding
NICTA, Australia
ming.ding@nicta.com.au

Lee Pike
Galois, Inc., USA
leepike@galois.com

Solar Energy Empowered 5G Cognitive Metro-Cellular Networks

Syed Ali Raza Zaidi, Asma Afzal, Maryam Hafeez, Mounir Ghogho, Desmond C. McLernon, and Ananthram Swami

ABSTRACT

Harvesting energy from natural (solar, wind, vibration, etc.) and synthesized (microwave power transfer) sources is envisioned as a key enabler for realizing green wireless networks. Energy efficient scheduling is one of the prime objectives in emerging cognitive radio platforms. To that end, in this article we present a comprehensive framework to characterize the performance of a cognitive metro-cellular network empowered by solar energy harvesting. The proposed model allows designers to capture both the spatial and temporal dynamics of the energy field and the mobile user traffic. A new definition for the "energy outage probability" metric, which characterizes the self-sustainable operation of the base stations under energy harvesting, is proposed, and the process for quantifying is described with the help of a case study for various UK cities. It is shown that the energy outage probability is strongly coupled with the path-loss exponent, required quality of service, and base station and user density. Moreover, the energy outage probability varies both on a daily and yearly basis depending on the solar geometry. It is observed that even in winter, BSs can run for three to six hours without any purchase of energy from the power grid by harvesting instantaneous energy.

Syed Ali Raza Zaidi, Asma Afzal, Maryam Hafeez, Mounir Ghogho, and Desmond C. McLernon are with the University of Leeds, UK.

Ananthram Swami is with the U.S. Army Research Lab, Adelphi, USA..

Mounir Ghogho is also affiliated with the International University of Rabat.

This work was supported by the U.S. Army Research Labortatory under Grant W911NF-13-1-0216.

INTRODUCTION

MOTIVATION

According to recent statistics [1], by the end of the year 2019 mobile broadband subscriptions are expected to reach 7.6 billion, accounting for 80 percent of all mobile subscriptions, compared to approximately 30 percent in 2013. Such an unprecedented incease in broadband demand will be further complemented by the exponential penetration of smart-phones, tablets, cyber-physical systems, machine-to-machine (M2M) communication devices, and mobile cloud based services. Gartner has predicted that Internet-of-Things (IoT) devices will grow to 26 billion units, representing a 30× growth compared to 2009. Similarly, Cisco Internet Business Solutions Group (IBSG) has forecasted that by the year 2020 the average number of Internet connected devices per person will amount to 6.8, compared to 1.8 in 2010, i.e. 50 billion Internet connected devices for the estimated world population of 7.6 billion. The steep ascent in demand inherently translates into a traffic explosion. The demand for mobile data traffic is expected to grow at a compound annual growth rate (CAGR) of 45 percent between 2013 and 2019. Consequently, it is predicted that while voice traffic will maintain its current trend, data traffic will grow 10 fold by the end of 2019 [1].

These formidable capacity demands have led to a so called "1000× mobile data challenge" introduced by Qualcomm. More specifically, the 1000× challenge dictates that fifth generation (5G) wireless networks, which are expected to roll out by early 2020, should be designed to be 1000 times more efficient than existing networks. In order to enable such a high level of efficiency, the architecture has to leverage three vital building blocks:

- Spectral agility.
- Network densification.
- Ultra energy efficient protocols.

While it is almost certain that the aforementioned architectural blocks should be combined in an efficient manner to address the so called "*exabyte flood*," the key question is how these blocks can be unified in a flexible architecture. Specifically, several design challenges for 5G networks are a byproduct of the trade-offs that exist in combining these architectural elements. The specific challenges addressed by each of these architectural pillars and the resulting trade-offs can be summarized as follows.

Pillar #1–Network Densification: As recognized in 3GPP LTE releases 10 and 12, network densification by small cell deployment plays an instrumental role in expanding wireless channel capacity. Intrinsically, the reduction in cell-size has a two-fold impact:

- Spatial load reduction, which is attained through both an increase in the degrees-of-freedom due to the multiplexing gain, and a reduction in the number of users per cell.
- Spectral aggregation, mainly due to the aggressive reuse of available transmission resources.

0163-6804/15/$25.00 © 2015 IEEE

While network densification is a promising solution to improve spectral efficiency, it must be complemented with an *interference coordination* (i.e. control, mitigation, and/or avoidance) mechanism to realize its full potential. A careful design is required, as implementation of such a mechanism has its own cost in terms of both the circuit and the transmit power consumption.

Pillar #2–Spectral Agility: It is well known that the sporadic utilization of available electromagnetic spectrum induces an artificial scarcity. The impact is more pronounced in the context of 5G wireless networks, where gains of 10–100× must be realized on top of data rates supported by legacy systems. The artificial spectrum scarcity can be mitigated by provisioning dynamic spectrum access (DSA) mechanisms. Most of the existing DSA mechanism aim to exploit one or more of these parameters in an opportunistic manner to enhance spectral efficiency. While opportunism enhances spectral utilization, the price paid is increased power consumption. In particular, the operational environment awareness is driven from the inference process which consumes more energy as compared to simple radio platforms.

Pillar #3–Energy Efficient Network and Protocol Design: The issue of so called "green design" is brought into play due to a predicted high volume of Internet connected devices. Specifically, as predicted in a recent report by Ericsson [2], the CO_2 emissions due to the number of Internet connected devices will increase from 800 Mtonnes to 1200 Mtonnes by 2020. Hence, like all other sectors, ICT should significantly reduce energy consumption to operate in an eco-friendly manner.

Is Cognitive Radio a Potential Solution?

Cognitive radios (CRs) are the key enablers for provisioning DSA. CRs are based on opportunistic exploitation of radio spectrum across one or more degrees of freedom. As observed in [3], CR-enabled DSA mechanisms can be alternatively considered as an interference management mechanism, i.e. these strategies effectively translate into interference control, shaping, and avoidance. In a nutshell, CR inspired small cellular networks are promising in terms of providing higher spectral efficiency due to tried and tested co-existence solutions. The state-of-the-art CRs naturally complement network densification by addressing the challenge posed in terms of interference coordination. Moreover, spectral agility is an intrinsic feature of the CR empowered network design. However, these intelligent radio terminals effectively trade energy efficiency for increased spectral efficiency. Consequently, this necessitates the design of next generation CRs that are capable of collectively addressing all previously stated design trade-offs. In other words, these desired design objectives serve as a blueprint for the requirement specification of the next generation of CRs.

The Second Generation CRs for 5G

Requirement Specification: For 5G wireless networks, CRs must take a leap forward in terms of opportunism providing gains both in terms of spectrum utilization and energy efficiency. Particularly, CR empowered small cells should:

- Maximize spectral efficiency by opportunistically utilizing the transmission vacancies across the spatio-temporal domain while co-existing in a heterogeneous network (HetNet) environment.
- Minimize energy consumption while opportunistically harvesting energy from ambient sources. The harvested energy will serve as a supplementary source to enable either a self sustainable eco-friendly operation or to accommodate an increasing number of users.
- Enable co-existence of small cell networks in a manner that is flexible, demand-adaptive, and self-organized. We need to enable co-existence through transmission, handover, and resource allocation coordination across different tiers and between different cells in the same tier. This may be provisioned in a distributed or centralized manner, depending on the overall architecture of network.

Small cells empowered by CRs that can combine the above mentioned attributes can be described as "second generation CRs." Contrary to the first generation CRs where exploitation of transmission opportunities was the prime objective, the second generation CRs will additionally be geared toward exploitation of energy harvesting opportunities from natural (solar, wind, vibration, etc.) and synthesized (microwave power transfer) sources.

Harvesting: Natural vs. Synthesized Sources: The key measure of the rate at which the power arrives on a unit area is termed "irradiance." Irradiance is the radiative flux measured in W/m^2. The amount of power harvested by employing a natural or synthesized source is an increasing function of irradiance experienced at the transducer. Generally, the input-output relationship of the transducer is non-linear. Thus, the output load is often matched to provide a maximum energy transfer.

Table 1 summarizes the typical values of irradiance observed at a transducer's input for various energy sources. As is clear from the table, solar and wind energy provide a minimum of 15× gain when compared with the next largest source, i.e. vibrational energy. It should be noticed that ambient RF energy has 10× lower irradiance than indoor solar irradiance. Consequently, harvesting from natural energy sources to empower self-sustainable small cellular networks seems a natural and plausible choice.

Problem Statement and Contributions

In order to explore the design space of the energy harvesting empowered CR small cellular networks, an adequate and meaningful metric is required. It is natural to assume that such a metric will be strongly coupled with both the dynamics of the energy harvester and the power consumption profile of a small cellular network. Furthermore, both of these factors are constructed by various important building blocks/parameters that jointly characterize the "*network-level self-sustainability*" (which will be defined in the subsequent discussion). Our main objective in this article is thus three fold:

- To highlight the key parameters that determine the dynamics of the harvester and shape the network-wide power consumption profile.

The amount of power harvested by employing a natural or synthesized source is an increasing function of irradiance experienced at the transducer. Generally, the input-output relationship of the transducer is non-linear. Thus, the output load is often matched to provide a maximum energy transfer.

Source	Irradiance	
Solar	Outdoor (solar noon)	100 mW/cm^2
	Outdoor (cloudy)	10 mW/cm^2
	Indoor	10–100 μW/cm^2
Wind	10 miles Class 7	12 mW/cm^2
Thermo-electric	5°C gradient	40 μW/cm^3
RF	ambient	1 μW/cm^2
Vibrations	Piezoelectric-shoe inserts	330μW/cm^3
	Electrostatic @ 105 Hz	0.021μW/mm^3
	Electromagnetic @ 10 Hz	184 μW/cm^3
	Electromagnetic @ 52 Hz	306 μW/cm^3

Table 1. Power densities of energy harvesting technologies.

- To present a new definition for a well known "energy outage probability" (EOP) metric to quantify the network-level self-sustainability. Notice that while the term EOP has been frequently used in recent harvesting literature, a statistical definition that can capture the specific dynamics for a natural energy harvesting empowered CR metro-cellular network is yet to be developed.
- To demonstrate the potential gains that can be harnessed by employing the proposed second generation CR enabled small cellular network deployments in a practical setup.

Empowering small cells with energy harvesting from a natural source such as the sun has also been solicited by Alcatel-Lucent in [4]. To this end, in this article we focus on the design space of solar energy harvesting empowered small cellular networks. As demonstrated in Fig. 1, the network level power consumption profile is a function of:

- Network architecture.
- Network wide load model.
- Desired quality of service (QoS) for mobile users (MUs).

Moreover, the dynamics of the harvester are coupled with:

- The spatio-temporal behavior of the ambient energy field.
- The properties of transducers that convert the ambient field into usable power.

Thus, to address the first objective, we highlight various design choices for the 5G metrocellular network which in turn determine the required operational power. We then identify the key parameters that dictate the characteristic of the solar energy field. We briefly outline the process of modeling the transducer, i.e. the photo-voltaic (PV) panels output in terms of the ambient input irradiance by considering an equivalent circuit model. The output power of the PV module is the key factor in characterizing the availability of energy to empower the operation of small cellular networks.

To address the second and third objectives, the solar energy harvesting model needs to be superimposed with a realistic spatio-temporal traffic and network model to characterize a network-wide performance metric. The metro-cellular networks are considered since small cell based densification by mounting platforms such as lightRadio® on the lamp posts is becoming increasingly common. We formulate the spatio-temporal model for mobile users (MUs) and metro-cellular base stations (BSs) from measurements obtained from different cities in the UK. With the help of a case study, we present:

- The modeling approach for capturing the behavior of the network load, the deployment topology, and the desired QoS for MUs.
- The gains exercised under the proposed deployment architecture.

Building Blocks for Cognitive Metro-Cellular Networks

There are several key design parameters which are crucial in characterizing the power requirements of the small cellular BS.

Deployment Mode: Metro cells can be deployed in a non co-channel or a co-channel mode. Metro cells can operate in a cognitive underlay mode where the same resource blocks are shared by the macro cells and small cells; power control at the small cells is implemented such that the MUs desired QoS requirements can always be guaranteed. An alternative phantom small cellular architecture is proposed by DOCOMO in [5, 6], where the metro cells are deployed in a non co-channel mode. Specifically, both the metro cells and the macro cells operate on different frequency bands. The architecture leverages the master-slave relationship between the macro cells and the metro cells, resulting in the separation of the control plane and the capacity plane (frequently known as the C/U plane split), thus providing support for adding capacity on-demand.

Deployment Location: Metro cells can be deployed uniformly across the macro-cellular network or alternatively on the cell edges to boost the capacity of the edge user. Even with a uniform deployment, the co-channel operation and the power control may push the operational region of the metro cells toward the edges as the interference aggregated from these edges may not deteriorate the performance of the users located toward the cell center.

Cloud vs. Traditional Radio Access Network (RAN): Cloud RAN (C-RAN) leverages its flexible architecture to provide coverage and capacity expansion in a cost efficient manner. Unlike traditional small cell networks, C-RAN architecture exploits the advantages of centralized baseband processing to address co-existence and scheduling issues. More precisely, C-RAN decouples the baseband processing unit (BBU) from the remote radio head (RRH). RRHs are connected to the cloud BBU pool via a flexible front-haul, which is usually a fiber optic cable where signaling is done using radio over fiber (RoF) or the common public radio interface (CPRI).The

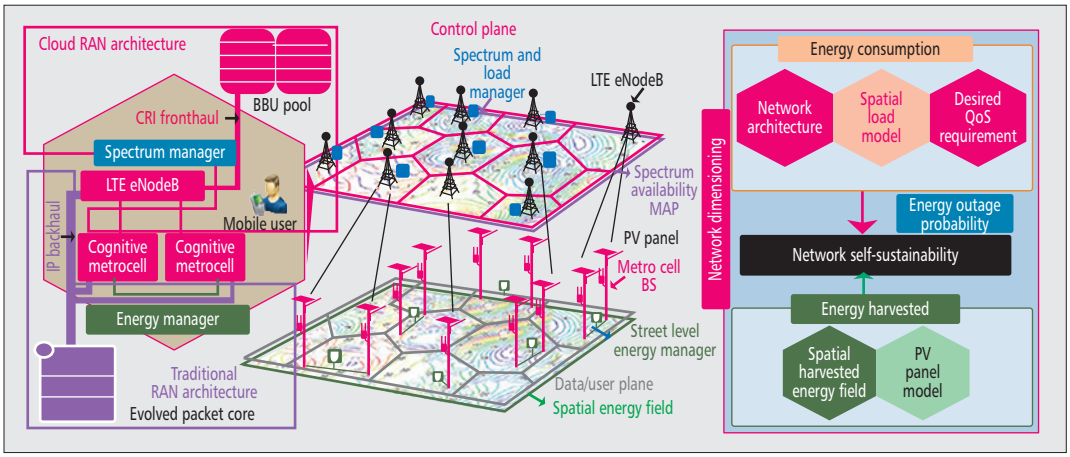


Figure 1. 5G cognitive metro-cellular network architecture.

amount of power required by the metro cell thus depends on whether the C-RAN architecture is implemented (where transmission and interference coordination provides significant gains in power reduction) or the traditional architecture is implemented.

Interference Coordination: The attainable performance of the network is mainly limited by the intra-tier and inter-tier interference. Realizing that interference is the key bottleneck, 3GPP LTE releases 10 and 11 have proposed enhanced inter-cell interference coordination (eICIC) schemes. In release 10, the concept of almost blank sub-frames (ABS) was introduced to elevate the downlink (DL) performance of a small cell user by scheduling it in the so called "blank sub-frame" of the macro cell. The concept was further extended to reduce power sub-frames (RPS) in LTE release 11 under the umbrella of further enhanced inter-cell interference coordination (FeICIC). Generally, the interference management techniques proposed under eICIC and FeICIC require either intra-tier and/or cross-tier coordination. Since the transmit power requirements are dictated by the QoS constraints expressed in the form of the desired signal-to-interference-plus-noise-ratio (SINR) threshold or throughput, the interference coordination strategy is also an important instrument in shaping the required transmit power of the metro cell.

The overall architecture of 5G cellular networks is depicted in Fig. 1. The characterization of the transmit power requirements with these architectural components will be demonstrated later with the help of a case study.

Dynamics of an Ambient Solar Energy Field and Modeling Output Power

Solar irradiance is an instantaneous measure of the energy arrival rate and thus varies across both the spatial and the temporal domains. Quantification of solar irradiance requires a comprehensive description of the underlying meteorological parameters. To this end, we provide a brief overview of these parameters. At this juncture it should be highlighted that in the recent past there has been enormous interest in studying cellular networks empowered by ambient RF energy harvesting. However, most of these studies assume stochastic/probabilistic energy arrival models. In practice, energy harvested from natural sources such as the sun have a significant deterministic component. Consequently, these models cannot accurately predict energy deficiency of power at a given time or day in a precise manner. Accurate prediction is of more interest to the cellular operator than an average performance metric.

Solar Insolation

The radiation intensity at the sun's surface is 6.33×10^7, $^W/_{m^2}$. The earth revolves around the sun in an elliptical orbit with the mean separation $r_{SE} = 1.496 \times 10^8$km (also known as 1 AU (astronomical unit)). Due to the distance squared spread of the radiant power, the amount of solar energy received outside the earth's atmosphere is reduced to $I_{SC} = 1367^W/_{m^2}$. The constant I_{SC} is frequently referred to as the 'solar constant'. The irradiance measured outside earth's atmosphere is generally called the extra-terrestrial (ET) solar irradiance. The energy that passes through the atmosphere and strikes the surface of a PV module is referred to as insolation. A number of astronomical and geometrical factors govern the amount of insolation, e.g. declination angle, zenith angle, latitude, longitude, day number, and atmospheric conditions [7]. Figure 2 provides a graphical illustration of these geo-physical parameters.

Using decades of past insolation and weather forecast data, numerous analytical models have been formulated taking all the above mentioned factors into account in order to characterize the direct and diffuse components of solar energy received by a PV module (see [7] for details). For the purpose of dimensioning a metro-cellular network, a simple yet accurate Hottell's clear day model [8] can be adapted to characterize the global horizontal irradiance in the absence of cloud cover. More sophisticated models can be employed to capture the randomness induced by cloud cover and aerosol absorption.

In practice, energy harvested from natural sources such as the sun have a significant deterministic component. Consequently, these models cannot accurately predict energy deficiency of power at a given time or day in a precise manner. Accurate prediction is of more interest to the cellular operator than an average performance metric.

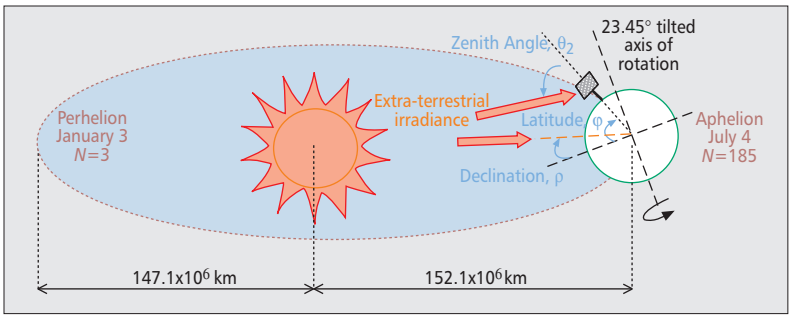


Figure 2. Illustration of the geophysical parameters controlling irradiance.

Harvested Solar Energy

Solar panels are comprised of PV cells made with various different materials such as mono-crystalline, polycrystalline, thin-film, or amorphous silicon. The underlying material determines the overall cost, panel efficiency, and power rating. In practice, the maximum output power is extracted by adjusting the cell load. The maximum extracted power is obtained by maximizing the output power with respect to the output voltage. Given the value of insolation at a particular time, the harvested power can be calculated using a well known single diode model for a PV module [9]. This computation requires a set of various parameters, such as: short circuit current; open circuit voltage; maximum power point voltage; and maximum power point current. These parameters can be individually expressed as a function of ambient temperature, insolation, and other constants that are specific to the panel itself and can be easily found in the panel's data sheet, such as voltage, current, and insolation values at standard temperature conditions (25°C). A detailed characterization of the output power (P_{PV}) in terms of these parameters is presented in [7]. The output power of the PV module (P_{PV}) can be compared with the power requirement of a metro-cellular BS to determine the network-wide self-sustainability.

Feasibility Study for Metro Cells Deployed on Lamp Posts in the UK

In this section we investigate the feasibility of deployment of solar energy powered metro-cellular networks in the UK. We build our case study by selecting worst-case features from the previously discussed network architecture for a power consumption profile. The relevant assumptions and considerations are presented in a subsequent discussion.

Assumptions and Considerations

Choice of Network Architecture and the PV Panel: We assume a two-tiered C/U plane split small-cellular network supported by the traditional IP back-hauling. Moreover, to simplify the analysis, we do not consider any intra-tier interference coordination mechanism. For the purpose of this case study, we focus on the DL operation of the considered cellular network. It is assumed that each CR metro-cellular BS is furnished with a PW1650-24V solar panel (as in [7]).

Network Deployment: Lamp posts can serve as ideal candidates for ultra dense outdoor BS deployment. To evaluate the effectiveness of metro cell deployment on lampposts, we obtained their coordinates from various cities in the UK including Nottingham, Winchester, Southampton, Basingstoke, and Salford. We also acquired the measured data for solar radiation from the British Atmospheric Data Centre (BADC). Our objective is to quantify the time for self-sustainable operation of metro cells with and without the presence of an energy storage device such as a battery. If the available energy is unable to satisfy the minimum rate requirements of the user, the network is said to be in energy outage. Thus, as implied by the name, EOP is the probability that the energy harvested from solar panels is not sufficient to fulfill consumer demands and thus additional energy has to be procured from the grid. Consumer demand can be adequately captured by the QoS parameters such as desired DL transmission rate and link reliability guarantees. In what follows, we present a relationship between the energy outage and the spatio-temporal dynamics of metro cells, user traffic, and the solar radiation for the above mentioned cities in the UK.

A Framework for Characterizing EOP

The quantification of power requirements for the CR metro-cellular network requires the spatio-temporal models for the MUs and the metro cells. It should be noted that the required power has a stochastic nature due to fading experienced on the communication links between an MU and its serving metro cell. The required transmit power is coupled with the desired data rate requirements (which forms the QoS constraint).

Spatial Model for Metro Cells: In the recent past, stochastic geometry has been used extensively for analyzing the performance of large scale cellular and ad hoc networks. Most of the studies assume that the spatial distribution of the BSs follows a homogeneous Poisson point process (HPPP). The key advantage of employing the HPPP based models is the analytical tractability of performance metrics such as coverage and ergodic rate. In [7] we used the nearest neighbor statistic for testing complete spatial randomness using the Clark-Evans test and showed that the spatial distribution of lamp posts in central Nottingham can be approximated by a HPPP. We conducted the same test for various cities in Hampshire county and the city of Salford. It was observed that the HPPP approximation also holds for these cities. The validity of HPPP based modeling can be intuitively explained by the fact that there is still sufficient randomness in deployment topology embedded due to the urban geometry even though the lamp posts exercise some degree of repulsion and regularity (as compared to the HPPP). Consequently, it is safe to capture the spatial configuration of the lamp posts by a HPPP with intensity λ_M for the purpose of this study.

Spatio-Temporal Model for MUs: In DL operation of the metro cells, the power required to serve MUs directly depends on the number of MUs in a cell. The number of active users changes with time, and their average density in a cell is known to vary according to a half sine

model with respect to the number of hours of the day. The sinusoidal variation of mean density has been derived from empirical traffic measurements collected from operational cellular networks. On a typical day user density has a predictable pattern, i.e. it reaches a minimum value around 4-5 a.m., rising steadily thereafter to a peak value in the evening and declining afterward. The distribution of the number of active users may change drastically for weekdays and weekends and also depends on the rate requirements.

Besides modeling temporal variations of MU, an accurate spatial distribution is also required to develop the load/activity model for small cellular networks. It has been demonstrated in the past that the active DL users are distributed according to a HPPP. Since the active users at a certain time in a cell are also the number of users distributed across space, the HPPP based modeling of MUs can be quite reasonable. Association of MUs with a serving metro cell can range from the nearest neighbor criterion to a more complex function of SINR in the presence of channel state information. For simplicity, we assume that an active user associates itself to the nearest BS, resulting in a Voronoi tessellation of metro-cellular BSs.

Quality of Service and Transmit Power Selection: The criteria for successful DL communication for CR small-cellular BSs is satisfied when a certain fraction of metro cell BSs in the network are able to meet the minimum rate demand for the active users. Mathematically, the probability of successful DL communication is given as

$$\mathbb{P}_{suc}^{MU} = \Pr\left\{ f\left(SINR\right) > R_o \right\} \geq \rho_{th}, \qquad (1)$$

where R_o is the MU's desired DL rate, ρ_{th} is the link reliability constraint, and f(SINR) is the instantaneous rate. For the purpose of this study, it is assumed that the minimum rate requirement is the same for all users. The minimum transmit power required (P_{MC}) to serve an MU such that its QoS requirements are satisfied can be established by quantifying P_{suc}^{MU} and then inverting the inequality in Eq. 1. The total power requirement to simultaneously serve all users in an arbitrary cell is $N_u(t)P_{MC}$, where $N_u(t)$ is the number of active DL users in an arbitrary metro cell.

Energy Outage Probability (EOP): When the metro cells operate without a battery, the user demand is met only when the instantaneous harvested energy is sufficient. In this case, the instantaneous EOP of the metro-cellular network is mathematically characterized as

$$\mathbb{P}_{out}^{E} = \Pr\left\{ N_u P_{MC} > P_{PV} \right\}. \qquad (2)$$

The above equation implies that the instantaneous EOP is distributed according to the spatial distribution of the number of active users in a cell. Network-wide self-sustainability can be characterized in terms of the average number of hours for which $\mathbb{P}_{out}^{E}$ is below a certain pre-specified threshold E_{out}. When the cognitive metro-cellular BSs employ P_{MC} to serve each active user, the average number of self-sustainable hours of operation per month are depicted in Fig. 3. As expected, during the summer the maximum harvested energy is significantly higher than in the winter. Consequently, the CR enabled metro-cellular network is self-sustainable for 10 to 12 hours on average during the summer for the considered set of parameters. Nevertheless, even during the winter approximately three to six hours of operation can be guaranteed by instantaneous expenditure of harvested solar power. The self-sustainable number of hours are also coupled with λ_U (user density). More specifically, it decreases with an increase in the mean number of peak hour users per cell. In this paper we consider $\lambda_U = \lambda_M N_s$ where $N_s = 6$ is considered.

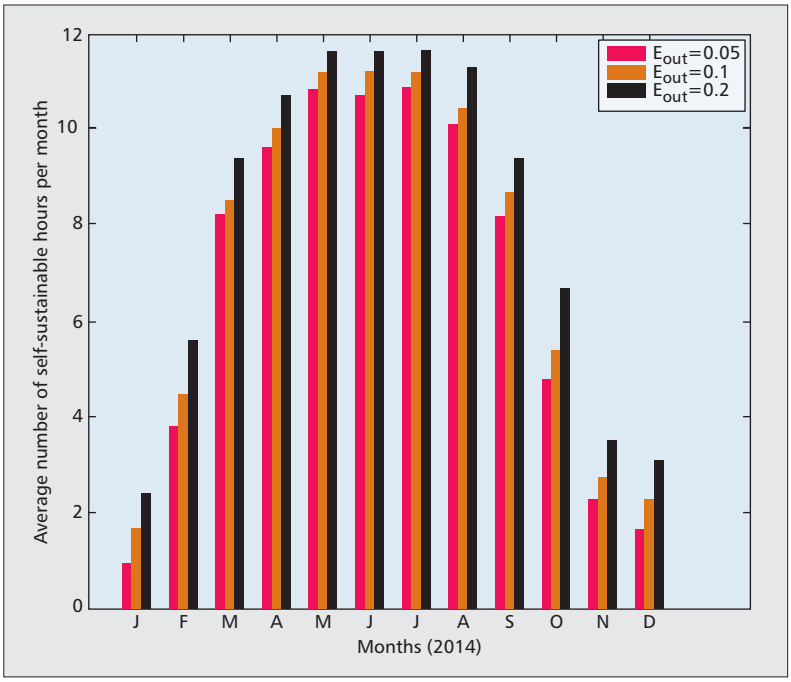


Figure 3. Energy outage probability of a metro-cellular network for λ_U (user density) = $6\lambda_M$ (metro-cell density), $\varepsilon = 0.02$, $\alpha = 4$, $\lambda_{LP} = 0.48 \times 10^{-3}$ (Nottingham City) and $\sigma^2 = -90$ dB/Hz.

With a practical solar energy storage system, performance can be improved significantly as the surplus energy during afternoons can be used in the evenings. Moreover, employing an appropriate sizing of the battery, continuous self-sustainability can be realized. An example of this would be to select a trickle charge battery to avoid the negative effects of over-charging in summers and have dimensions such that a metro cell can utilize battery reserve for a number of consecutive overcast days. We conducted a simplified analysis to quantify the energy deficit when the metro cell BS is equipped with a battery. We considered a 12v, 1 ampere hour battery attached to a metro cell BS and compared the average daily energy demand with the average daily harvested energy, each multiplied with the number of days in a particular month. Neglecting the charging and discharging inefficiencies, the average battery state for a particular month is simply considered to be the difference between the demand and supply, as shown in Fig. 4. The results in Fig. 4 show that a battery operated metro-cellular network may never go into outage for the entire year except in the month of January.

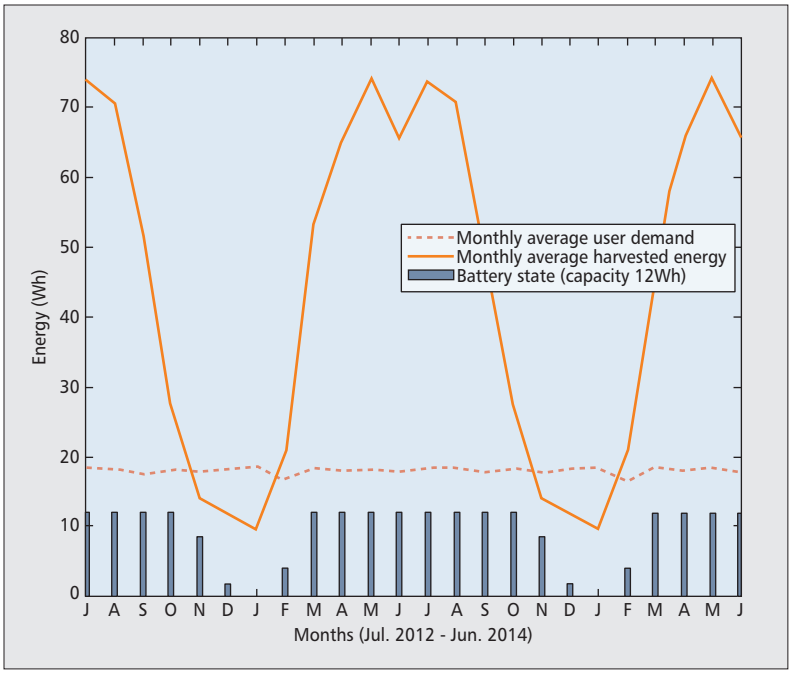


Figure 4. Average daily harvested energy and user demand aggregated over a month and the corresponding average battery state. λ_U (user density) = $6\lambda_M$ (metro-cell density), $\alpha = 4$, $\lambda_{LP} = 0.48 \times 10^{-3}$ (Nottingham City) and $\sigma^2 = -90$ dB/Hz.

Open Issues and Future Directions

There are several important open design issues that need to be considered to explore the design space and full potential of CR empowered metro-cellular networks. Due to space limitation, we only highlight the most important and promising research directions.

Energy Storage: Dimensioning of energy storage, design of online prediction algorithms, and optimal trading of energy with the grid still remain open issues. With the advent of smarter grid infrastructure, energy trading and sharing on the micro level for self-sustainable network deployment will become possible in the near future. However, the impact of such an evolution when communication networks are empowered with green renewable sources has not yet been fully investigated.

Energy Aware Load Balancing: Tiered structures with heterogeneous small cells, relays, and distributed antenna systems are envisioned to be the key enabler toward addressing the 1000× challenge. In HetNet deployments, the transmit power of each child tier is generally lower than its parent tier. Thus, received signal strength based association may overload the parent tier due to higher transmit power. In order to circumvent this problem, 3GPP standards have introduced the concept of "biased-association." In particular, a range extension bias (REB) is introduced in received signal strength to determine the tier that will serve the MU. Due to the introduction of such bias, the MUs may experience high interference as they may be associated with a sub-optimal serving BS. However, performance can be elevated through inter-tier interference coordination. We advocate that for a 5G cognitive metro-cellular network, load balancing should be complemented with energy balancing, i.e. REB should designed such that the energy burden can be shared across the network. Specifically, due to the highlighted variations in both energy field and user traffic across space and time, association criterion should be adaptive.

Energy Aware Interference Coordination: As mentioned earlier, inter-tier and intra-tier interference coordination are important features of small cellular networks. As discussed earlier, EICIC or FeICIC are employed to manage inter-tier interference when co-channel deployment is the preferred option. The duty cycle of almost blank subframes or reduced power subframes is designed to optimize the throughput. However, in an energy harvesting empowered network, we suggest that the design of duty cycles should consider the natural variations in energy states. BSs with lower residual energy may increase blank subframe duty cycle to reduce co-channel interference while harvesting energy. The duty cycle should be optimized for throughput whenever sufficient energy is available at BSs to schedule transmissions. Optimal duty cycling for interference coordination in an energy aware manner has not yet been explored.

Conclusion

In this paper we presented a framework to investigate the performance of a solar empowered metro-cellular network. We depart from the traditional definition of cognition which focuses on spectral efficiency performance; rather, we characterized cognition in terms of energy efficiency. It is demonstrated that both temporal and spatial dynamics of the solar energy field and mobile user traffic are critical in shaping the network-wide energy requirement. The energy demand of a metro-cellular BS is also strongly coupled with the QoS desired by MUs. It is shown that a metro-cellular network is self-sustainable in terms of energy for approximately three to 12 hours of a day depending upon the time of the year. Finally, it was argued that the dynamics and randomness in energy state can be exploited in the future to attain energy aware load balancing and interference coordination.

References

[1] A. Ericsson, "Traffic and Market Data Report — On the Pulse of the Networked Society," 2014.
[2] —, "Ericsson Energy and Carbon Report: On the Impact of Networked Society," 2013.
[3] S. Zaidi, D. McLernon, and M. Ghogho, "Breaking the Area Spectral Efficiency Wall in Cognitive Underlay Networks," *IEEE JSAC*, vol. 32, no. 11, Nov. 2014, pp. 2205–21.
[4] A. Lucent, "Alternative Renewable Power Concepts for Alcatel-Lucent Outdoor Wireless Small Cell Solutions," 2011.
[5] H. Ishii, Y. Kishiyama, and H. Takahashi, "A Novel Architecture for LTE-B: Cplane/u-Plane Split and Phantom Cell Concept," *IEEE Globecom Wksps.*, 2012, pp. 624–30.
[6] S. Mukherjee and H. Ishii, "Energy Efficiency in the Phantom Cell Enhanced Local Area Architecture," *IEEE Wireless Commun. and Networking Conf. (WCNC)*, 2013, pp. 1267–72.
[7] S. Zaidi *et al.*, "Energy Harvesting Empowered Cognitive Metro-Cellular Networks," *IEEE 1st Int'l. Wksp. Cognitive Cellular Systems (CCS)*, 2014, pp. 1–5.

[8] W. B. Stine and M. Geyer, Power from the Sun. powerfromthesun.net, 2001.
[9] A. Bellini *et al.*, "Simplified Model of A Photovoltaic Module," *IEEE Applied Electronics*, 2009, pp. 47–51.

Biographies

Syed Ali Raza Zaidi (s.a.zaidi@leeds.ac.uk) is currently a research fellow at the University of Leeds on the U.S. Army Research Lab funded project, "Cognitive Green Wireless Communication–A Network Science Perspective." He received his B.Eng degree in information and communication system engineering from the School of Electronics and Electrical Engineering, NUST, Pakistan in 2008, and a Ph.D. from the University of Leeds in 2013. He was awarded the University of Leeds F. W. Carter prize for best Ph.D. thesis, and NUST's most prestigious Rector's gold medal for his final year project. From September 2007 to August 2008 he served as a research assistant in the Wireless Sensor Network Lab on a collaborative research project between NUST, Pakistan and Ajou University, South Korea. In 2008 he was awarded an overseas research student (ORS) scholarship along with Tetley Lupton and Excellence Scholarships to pursue his Ph.D. at the School of Electronics and Electrical Engineering, University of Leeds, U.K. He was also awarded with COST IC0902, DAAD, and Royal Academy of Engineering grants to promote his research. He was a visiting research scientist at the Qatar Innovations and Mobility Center from October to December 2013. He has served as workshop chair for IEEE VTC 2015 DCS, IEEE IWCMC 2015 ReAP, CROWNCOM 2015 and IEEE CAMAD 2015 ReAP. He has served as an invited reviewer for IEEE flagship journals and conferences. Over the past few years he has served as TPC member for IEEE WCNC 12-15, IEEE VTC 10-15, EUSIPCO 11-15 and SPAWC 10. He is also the lead guest editor for the *IET Signal Processing Journal* SI on signal processing advances for 5G. His research is focused toward the design and analysis of large scale ad-hoc wireless networks by employing tools from stochastic geometry and random graph theory.

Asma Afzal (elaaf@leeds.ac.uk) received the B.E. and M.S. degrees in electrical engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2011 and 2013, respectively. She is currently working toward the Ph.D. degree in electrical engineering from the University of Leeds, Leeds, U.K. During her masters she worked as a lab engineer at NUST on a collaborative project with Cypress Semi-Conductor, San Jose, CA, USA. Her research interests are in stochastic modeling and analysis of future generation wireless networks with energy harvesting.

Maryam Hafeez (elmh@leeds.ac.uk) received a B.Eng. in information and communications systems engineering from the School of Electrical Engineering and Computer Sciences (SEECS), National University of Science and Technology (NUST), Pakistan in 2008. Throughout her undergraduate degree she was awarded the SEECS prestigious merit scholarship. NUST's most presitigous Rector's gold medal was awarded to her final year project. From 2008 to 2009 she served as a research assistant in the Wireless Sensor Network (WiSNET) Lab on a collabrative research project between Ajou University, South Korea and the NUST, Pakistan. Her research is geared toward design and analysis of protocols for next generation green intelligent wireless networks by employing tools from game theory and stochastic geometry

Mounir Ghogho (m.ghogho@leeds.ac.uk) received the M.Sc. degree in 1993 and the Ph.D. degree in 1997 from the National Polytechnic Institute of Toulouse, France. He was an EPSRC research fellow with the University of Strathclyde, Glasgow (Scotland), from September 1997 to November 2001. He joined the school of Electronic and Electrical Engineering at the University of Leeds (UK) in December 2001, where he currently holds a chair in signal processing and communications. He is also currently a research director at the International University of Rabat (Morocco). He is currently an associate editor of *IEEE Signal Processing Magazine*. He served as an associate editor of *IEEE Transactions on Signal Processing* from 2005 to 2008, *IEEE Signal Processing Letters* from 2001 to 2004, and the Elsevier *Digital Signal Processing* journal from 2011 to 2012. He is currently a member of the IEEE Signal Processing Society SAM Technical Committee. He served as a member of the IEEE Signal Processing Society SPCOM Technical Committee from 2005 to 2010, and as a member of the IEEE Signal Processing Society SPTM Technical Committee from 2006 to 2011. He was the general chair of the 11th IEEE Workshop on Signal Processing for Advanced Wireless Communications (SPAWC'2010), general chair of the 21st edition of the European Signal Processing Conference (EUSIPCO 2013), the technical co-chair of the MIMO symposium at IWCMC 2007 and IWCMC 2008, and a technical area co-chair of EUSIPCO 2008, EUSIPCO 2009, and ISCCSP'05. His research interests are in signal processing and communication networks. He has published three book chapters, more than 75 journal papers and 150 conference papers. He is the EURASIP liaison in Morocco. He was awarded the prestigious and highly competitive five-year UK Royal Academy of Engineering Research Fellowship in September 2000. He is also one of the recipients of the internationally competitive 2013 IBM Faculty award. He has held invited scientist/professor positions at many institutions, including the US Army Research Lab (USA), TÉLÈcom Paris-Tech (France), the National Institute of Informatics (Japan), the University Carlos Third of Madrid (Spain), ENSICA (France), the Technical University of Darmstadt (Germany), the University of Minnesota (USA), Beijing University of Posts and Telecommunication (China), and the University Mohamed V (Morocco).

Desmond C. McLernon (d.c.mclernon@leeds.ac.uk) received the B.Sc. in electronic and electrical engineering and the M.Sc. in electronics from Queen's University of Belfast, N. Ireland. He then worked on the research and development of radar systems with Ferranti Ltd. in Edinburgh, Scotland, and later joined Imperial College, University of London to obtain the Ph.D. degree in signal processing. After first lecturing at South Bank University, London, UK, he moved to the School of Electronic and Electrical Engineering at the University of Leeds, UK, where he is currently a reader in signal processing. His research interests are broadly within the domain of signal processing for wireless communications systems (in which discipline he has published more than 270 international journal and conference papers). He is the associate editor of the UK IET journal *Signal Processing*. He has been a member of various international conference organizing committees, more recently the IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2010, Marrakech), European Signal Processing Conference (EUSIPCO) 2013, IET Conference on Intelligent Signal Processing (London, 2013 and 2015), and IEEE Globecom 2014 and 2015, Workshop on Trusted Communications with Physical Layer Security (Austin, Texas).

Ananthram Swami (a.swami@ieee.org) received the B.Tech. degree from the Indian Institute of Technology (IIT), Bombay, the M.S. degree from Rice University, Houston, TX, and the Ph.D. degree from the University of Southern California (USC), Los Angeles, all in electrical engineering. He has held positions with Unocal Corporation, USC, CS-3, and Malgudi Systems. He was a statistical consultant to the California Lottery, developed a Matlab-based toolbox for non-Gaussian signal processing, and has held visiting faculty positions at INP, Toulouse, France. He is with the U.S. Army Research Laboratory (ARL), where he is the ST for Network Science. His work is in the broad area of network science, with an emphasis on wireless communication networks. He was the co-editor of *Wireless Sensor Networks: Signal Processing & Communications Perspectives* (New York: Wiley, 2007). He is a member of the IEEE SPS Technical Committee on Sensor Array and Multi-Channel systems, and serves on the Senior Editorial Board of the *IEEE Journal on Selected Topics in Signal Processing*. He is an ARL Fellow. He has served as an associate editor for *IEEE Transactions on Signal Processing*, *IEEE Signal Processing Letters*, *Signal Processing Magazine*, *IEEE Transactions on Circuits and Systems II*, *IEEE Transactions on Wireless Communications*, and as guest editor for the *IEEE Journal on Selected Areas in Communications*. He was a tutorial speaker on "Networking Cognitive Radios for Dynamic Spectrum Access" at ICASSP 2008, DySpan 2008, MILCOM 2008, and ICC 2010. He received the best conference paper award at IEEE Trustcom 2008, and was co-Organizer and co-Chair of three IEEE workshops related to signal processing and communications, including IEEE SPAWC'10.

We advocate that for a 5G cognitive metro-cellular network, load balancing should be complemented with energy balancing, i.e. REB should designed such that the energy burden can be shared across the network. Specifically, due to the highlighted variations in both energy field and user traffic across space and time, association criterion should be adaptive.

Virtualized Cognitive Network Architecture for 5G Cellular Networks

Hesham ElSawy, Hayssam Dahrouj, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouini

ABSTRACT

Cellular networks have preserved an application agnostic and base station (BS) centric architecture[1] for decades. Network functionalities (e.g. user association) are decided and performed regardless of the underlying application (e.g. automation, tactile Internet, online gaming, multimedia). Such an ossified architecture imposes several hurdles against achieving the ambitious metrics of next generation cellular systems. This article first highlights the features and drawbacks of such architectural ossification. Then the article proposes a virtualized and cognitive network architecture, wherein network functionalities are implemented via software instances in the cloud, and the underlying architecture can adapt to the application of interest as well as to changes in channels and traffic conditions. The adaptation is done in terms of the network topology by manipulating connectivities and steering traffic via different paths, so as to attain the applications' requirements and network design objectives. The article presents cognitive strategies to implement some of the classical network functionalities, along with their related implementation challenges. The article further presents a case study illustrating the performance improvement of the proposed architecture as compared to conventional cellular networks, both in terms of outage probability and handover rate.

INTRODUCTION

The fifth generation (5G) cellular networks are expected to offer ubiquitous and global connectivity for everything (users, devices, sensors, machines) and support diverse types of applications with different operational constraints. Some of these applications are delay sensitive (e.g. automation, tactile Internet), others are bandwidth (BW) aggressive (e.g. online gaming, multimedia), and others are demanding in terms of the numbers of connections (e.g. smart cities). Hence, context awareness (i.e. awareness of the application requirements and real-time network related information such as the network conditions, source and destination relative locations, interference levels, congestion bottlenecks) is crucial for 5G networks to effectively support these diverse applications. However, this is not the case for conventional cellular networks, which have historically preserved an application agnostic and base station (BS) centric architecture. Regardless of the underlying application, network conditions, and relative locations of the devices, the network functionalities (e.g. user association, resource allocation, data routing etc.) are performed in the same manner. This may lead to inefficient resource utilization in the radio access network (RAN) and unnecessary delays in the core network. Hence, such a problem, denoted in this article by *architectural ossification*[2], spans both the access and core networks. This article focuses on the RAN problems and discusses the potential solutions.

Architectural ossification may also impede the evolution toward the ambitious metrics defined for the 5G networks, namely, the 1000-fold capacity increase with at least 100-fold leap in the peak data rate and 0.1x delay reduction [1]. In particular, network densification via small cells and millimeter wavelength (mmW) communication, which are the main drivers for capacity and data rate improvement, require a flexible network architecture. For instance, the migration to the mmW band may impose spatial blind spots to the BSs' coverage, due to the significant effect of shadowing on mmW propagation, which requires redundant associations to increase the probability of LoS connection (e.g., serving one user by several BSs). Also, conventional association in dense small-cell environments imposes considerable handover signaling due to mobility, hence, new association schemes are required for handover signaling reduction.

To obtain the desired 5G performance metrics, cognition and flexible architecture realized via context aware network functionalities becomes a necessity. This article integrates recent advances in cellular networking and proposes a unified virtualized and cognitive architecture wherein the control and data planes are decoupled under a cloud-RAN (CRAN) umbrella, so as to adapt the network functionalities to changes in channels and traffic conditions as well as to the underlying application. Decoupling the control plane and the data plane, which is proposed in [2], enables centralized soft-

Hesham ElSawy, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouni are with Abdullah University of Science and Technology (KAUST). Tareq Y. Al-Naffouri is also with King Fahd University of Petroleum and Minerals.

Hayssam Dahrouj is with Effat University.

Network architecture, in this article, defines how BSs are connected to each other and how the cellular network elements (e.g., BSs, user, relays, etc.) communicate.

2 We borrow the terminology form the Internet ossification problem due to the analogy between the two cases.

0163-6804/15/$25.00 © 2015 IEEE

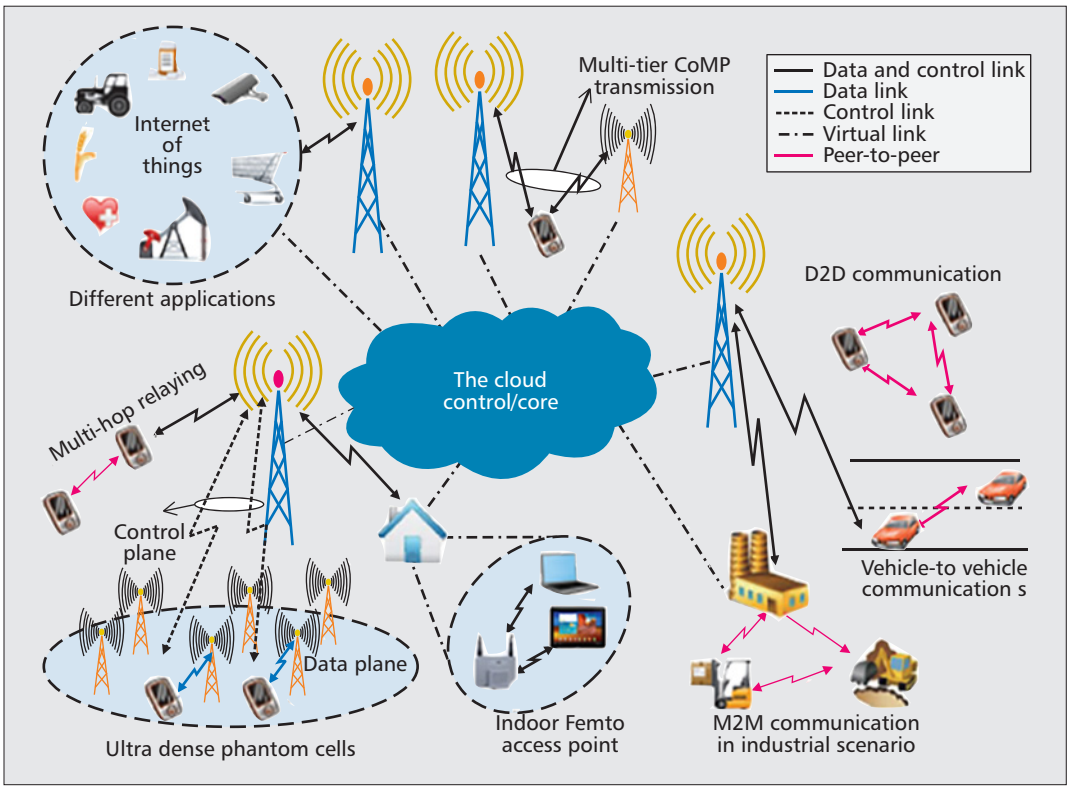


Figure 1. Schematic digram for 5G cellular network. The cloud is the core engine of the network which monitors the traffic spatial and temporal variation as well as the traffic classes, based on the underlying application, and controls the network operation accordingly.

It is worth noting that centralized network control in the cloud does not necessarily mean a centralized execution for network functionalities. Some network functions may be implemented in a distributed cognitive manner while the cloud only dictates the operational guidelines.

ware control for network behavior, i.e. instead of using certain modules to perform network functions at each and every BS, a single script written in the cloud can control the entire network behavior. The forwarding plane, however, remains distributed according to the physical locations of the network entities (i.e. BSs, servers, gateways). This leads to a self-organizing network (SON), and provides generous flexibility both in terms of network expansion via BS deployment, and in terms of creating new services and applications.

It is worth noting that centralized network control in the cloud does not necessarily mean a centralized execution for network functionalities. As will be discussed later, some network functions may be implemented in a distributed cognitive manner while the cloud only dictates the operational guidelines. For instance, a proximity application using device-to-device communication can opportunistically access the cellular channel, in which the interference constraints for the cellular users are defined by the cloud.[3] In such a *virtualized, hybrid centralized and distributed architecture*, there is no single rule to execute network functions. Instead, network functions are performed based on the network condition and underlying application. According to the traffic requirements and bottleneck location, the cognitive cloud would steer the traffic from one path to another, or replace a single hop congested BS link with a multi-hop non-congested link to offer adequate quality-of-service (QoS) and meet the application constraints. For instance, an emergency automation message targeting an actuator in the proximity area around the sensor does not have to go through the serving BS and core network. Instead, it can be directly conveyed to the actuator in a multi-hop fashion. Further, the cloud may suppress co-channel interference by nearby devices and BSs during the automation message transmission, so as to ensure an error free delivery for the emergency message.

CRAN is also an enabler for proactive, instead of reactive, network control. The cloud can gather information (e.g. from social networks) and learn about users' interests, preferences, data usage, and mobility patterns. This information can be used for proactive resource allocation and data caching, which can substantially enhance network performance [3]. Figure 1 illustrates the proposed architecture and emphasizes the different supported services. The figure shows the different network connectivity types (i.e. multi-hop, device-to-device, control signaling decoupling) in which the control engine is located at the cloud. Context awareness at the cloud enables the network to respond differently to different applications and network conditions.

The remainder of the article presents practical strategies of implementing cognitive network functionalities. Then it introduces the related implementation issues. The article finally shows potential gains of the proposed virtualized cognitive architecture through illustrative simulations.

Network Functionalities

From the RAN perspective, network topology is defined by the connections between the network entities. Figure 2a shows the star-shaped topolo-

[3] It is important to highlight that estimating the channel gains towards the receivers, which is infeasible in TV white space cognitive network due to their passive TV set receivers, may be feasible in cellular networks. This is because cellular networks are characterized by their active elements, which simultaneously transmit and receive data. Hence, the channel can be estimated during the transmission period.

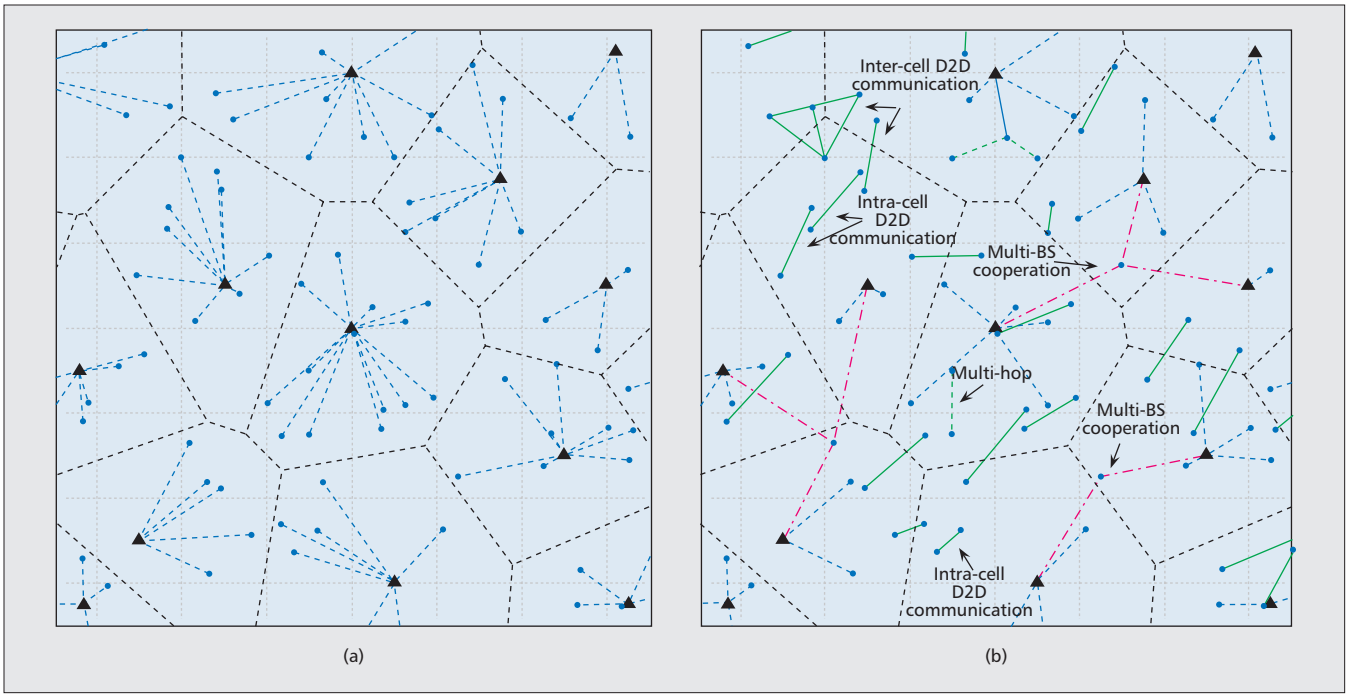


Figure 2. Network topologies for the same locations of BSs and UEs in which the triangles represent the BSs and the dots represent users, black dotted lines represent cell boundaries, blue dotted lines represent single BS connectivity, red dotted lines represents multi-BS connectivity, and greed dotted lines represent peer-to-peer D2D connectivity: a) BS centric topology (*conventional star-topology*) where each UE connects to its nearest BS; and b) context aware topology in which the connections are established based on several aspects such as the relative distance between nodes, application, SINR, …etc.

gy enforced by conventional cellular networks. The star-topology emerges from the fact that all transmissions are routed through BSs, regardless of the underlying application and the required data flow, in a centralized manner to guarantee efficient interference management. The communication within the conventional cellular network is merely "*BS centric*" where cell boundaries determine the region served by each BS. The relative radio signal strength (RSS) between neighboring BSs is the common way to determine cell boundaries. This section revisits some of the classical network functionalities and shows how redefining such functionalities in a cognitive fashion yields an appreciable flexible network operation, as shown in Fig. 2b.

User Association

Conventional Operation: User association is the most basic, yet critical, network function in cellular networks, as it assigns users to BSs. User association controls the traffic and load served by each BS and each network tier. Association impacts many performance metrics such as coverage probability, rate, blockage, etc. Therefore, many attempts have been made to optimize BS association in cellular networks [4, 5]. The common message in these works is that RSS based association is not always efficient. For instance, in [4] the authors show that different users' association strategies are required to attain different design objectives (e.g. delay, rate, fairness) in a single tier cellular network. In [5] it is shown that the inefficiency of RSS based association is more prominent in multi-tier networks due to the high transmit power disparity between BS types (i.e. macro, micro, pico, and femto). In fact, the authors show that with the proper manipulation of the association function, in a two tier cellular network the minimum rate attained by the users can be increased by four times.

Proposed Operation: The virtualized network architecture not only renders the RSS rule obsolete and allows a flexible cell association, but also allows decoupled uplink, downlink, and control association. To illustrate how virtualization achieves such decoupling, consider a three-tier network in which a test user has a pico-BS as their closest BS, then a micro-BS farther than the pico-BS, and then a macro-BS farther than the micro-BSs. However, due to the downlink transmit power disparity, the macro-BS (pico-BS) provides the highest (lowest) downlink RSS. Instead of associating with the macro-BS only, which might be congested, the user can communicate in the downlink with the micro-BS for load balancing, and in the uplink with the pico-BS for transmit power reduction. To reduce the handovers caused by mobility, the user can receive control signaling from the macro-BS. Cognition, in this case, becomes important, since there is no single rule for association, as it depends on the underlying application and the network conditions. For example, if an application has tight rate constraint, an uplink connection to a less loaded, although much farther and may require higher uplink transmit power, BS may be more efficient than a congested nearby BS. Further, users' association has to adapt to the traffic and spatial distributions in order to attain the desired network objective and applica-

tion requirements. As shown later, significant gains can be harvested from such a flexible association.

Device-to-Device Communication

Conventional Operation: The conventional cellular infrastructure dictates that all traffic is communicated through the BSs, regardless of the users' relative locations (i.e. the required data flow). However, with the proliferation of proximity based services and social networking, nearby users may wish to establish connections and exchange data. Recent studies show that if nearby devices are allowed to bypass the cellular infrastructure and directly communicate in a peer-to-peer fashion, which is referred to as device-to-device (D2D) communication, many performance metrics can be improved [6]. D2D communication has the potential to offload traffic from congested cellular BSs and spatially reuse short-distance small-power peer-to-peer links. Recent studies [6] suggest that D2D communication can improve cellular system throughput by 374 percent, power efficiency by 100 percent, and cell edge users' performance by 300 percent, when compared to conventional star-topology. It is worth mentioning that D2D communication is the key enabler for the cellular network to support machine-to-machine (M2M) communication and Internet of Things. D2D communication enables a massive number of machines to communicate together and connect to the Internet. Admitting all of the machines' traffic to the cellular infrastructure overloads the network and results in congestion and blockage. Offloading such machine-type communication to the D2D mode whenever possible alleviates congestions and enhances overall network performance.

Proposed Operation: The proposed virtualized network architecture provides a flexible paradigm for enabling D2D connectivity. That is, on top of the decoupled uplink, downlink, and control signaling connectivity, there is an option to establish one or more of these connections in the D2D mode. For instance, consider two low power devices A and B such that A is located between B and a nearby BS. If A is closer to the BS than to B, the transmission from A to B can be established via the BS and the transmission for B to A can be established in the D2D, and hence, the transmit power is reduced. Such a flexible communication paradigm enables fine tuned traffic control to reduce congestions, reduce power consumption, and increase network efficiency. However, D2D communication increases the complexity of the admission process as it defines a new network function. That is, the admission process of each user includes a mode selection function to determine the mode of operation of each link, namely D2D mode or cellular mode. D2D communication also changes the conventional cellular star-topology to a hybrid topology (i.e. coexisting star and ad hoc topologies), in which interference management is the core challenge. In this case, cognition is important for interference management between D2D and cellular links. Note that the priority of the D2D and cellular links to use the spectrum is based on the application. For instance, as discussed earlier, a critical D2D automation message would be the primary spectrum user and the cellular links would be the secondary spectrum users. On the other hand, a file transfer via D2D mode would be the secondary spectrum user and the cellular links would be the primary users. Also, cognitive coordination between different D2D links for spectrum access is important to maintain an acceptable QoS.

Multicell Coordination

Conventional Operation: Conventional network architecture enforces one user to one BS association strategy. In this case, employing aggressive frequency reuse schemes, due to the scarcity of the wireless spectrum, imposes high inter-cell interference. Such interference is a main performance limiting parameter for cellular networks, especially for cell edge users. Multi-cell cooperation, via coordinated multi-point transmission (CoMP), is employed to reduce intercell-interference and improve the signal-to-interference-plus-noise-ratio (SINR) statistics. That is, multiple BSs cooperate (e.g. via beamforming) to simultaneously serve multiple common users. Such cooperation boosts the system capacity by reducing inter-cell interference and improving the SINR statistics.[4]

Proposed Operation: The virtualized network architecture enables multiple BS association for the uplink, downlink, and control signaling for each user. CoMP changes the network topology to mitigate inter-cell interference and enhance cell edge performance, as shown in Fig. 2b. However, this may increase the signaling overhead between BSs. In this case, a per user cognitive enabled/disabled CoMP scheme can be implemented. This leads to a flexible CoMP operation that accounts to the user SINR, the application requirement, and the state of the surrounding BSs (e.g. congested or not). Note that in a CRAN environment, BSs act as a virtual antenna system for the cloud, which transforms CoMP scheduling into a multiuser (MU) MIMO scheme.

Multi-Hop Relaying

Conventional Operation: Earlier, we discussed the D2D mode of communication as an alternative to the cellular link. A potential application for the D2D mode is multi-hop relaying. That is, a cellular link can be replaced by several consecutive D2D links. Multi-hop relaying is essential with the drastic increase in the population of devices. There could be situations where a massive number of machines simultaneously need to connect to the cellular infrastructure. One example is traffic jams where all smart vehicles need to connect to the Internet to send information or receive updates about the traffic conditions. Another instance may occur in over crowded places such as stadiums where a huge population of users want to simultaneously connect to the network. In these cases, the direct link connectivity imposed by the conventional cellular infrastructure results in high blockage probability and degraded user experience. Replacing the single direct link per channel per cell by multiple short-distance low-power links that can be reused several times within the same

The proposed virtualized network architecture provides a flexible paradigm for enabling D2D connectivity. That is, on top of the decoupled uplink, downlink, and control signaling connectivity, there is an option to establish one or more of these connections in the D2D mode.

4 Several cellular operators have already launched CoMP trials, also denoted as elastic cell, and demonstrated the performance improvement.

For feasible and efficient network operation, we seek tradeoffs between complexity, signaling, and performance. We advocate to split the network functions into two main categories, namely, instantaneously and statistically optimized functions.

cell area alleviate such congestion. It is shown in [7] that the appropriate design of multi-hop relaying can increase the minimum achievable rate of users by up to 100 fold.

Proposed Operation: The virtualized network architecture allows multi-hop relaying in each of the decoupled links. This can be exploited to extend BS coverage, where users in coverage holes can relay information to the BS via another covered user, or in high connectivity demand periods. An important application for multi-hopping is proximity-based services with delay constraints. In some cases the BSs are located hundreds of kilometers away from the nearest serving gateway (SGW) [8]. Hence, routing information in the conventional way via the core network (i.e. SGW) may encounter unnecessarily high delays. Therefore, multi-hop relaying is a key solution to satisfy the delay constraint for such applications. Multi-hop relaying defines a new network function that selects whether to multi-hop or directly send to the BS. Hence, multi-hopping changes the star-topology into an extended star-topology, which imposes an inherent routing problem. The cloud should be able to detect network congestion, in terms of users' population, and enable multi-hopping to maintain an acceptable blocking probability. As discussed earlier, the priority for the channel access between cellular and multi-hop D2D links highly depends on the underlying application.

Implementation Issues

The massive number of network elements (i.e. BSs, users, machines, etc.) makes a centralized instantaneous optimization for the network functions infeasible, even with super computing agents in the cloud. That is, it is infeasible to select a serving BS, assign powers, allocate channels, and choose the mode of operation for each and every network element within the cellular network. In this section we discuss some trends to compromise between complexity and performance in prospective cognitive CRAN networks, as well as the limitations for CRAN operation.

Feasibility and Complexity Tradeoffs

For feasible and efficient network operation, we seek tradeoffs between complexity, signaling, and performance. We advocate to split the network functions into two main categories: instantaneously and statistically optimized functions. While instantaneous optimization guarantees the best performance at any time instant, statistical optimization provides optimal averaged performance on a long-time scale to reduce signaling and processing overheads. Specifically, instead of requiring instantaneous information, which is difficult to obtain and communicate, statistical network parameters can be exploited to guarantee an average optimal performance. For instance, in [9], instead of optimizing the transmit power based on the instantaneous channel gains, a simple power policy, which is optimal to the channel gain distribution, is developed. It is worth mentioning that many attempts have been made to either statistically or instantaneously optimize network functions. However, to the best of the authors' knowledge, merging statistical and instantaneous optimization to balance performance, complexity, and signaling overhead has been ignored.

In addition to statistical optimization, cognitive and distributed control for some network functions can be exploited to reduce complexity. In this case, the network elements can choose their instantaneous operating parameters (e.g. power level, channel, and mode of operation) in a cognitive way to maximize the network objective subject to the enforced operator policies. For example, based on the relative D2D and cellular link distances, users can distributively select their mode of operation. Then, on one hand, D2D users can operate in a distributed and cognitive manner. On the other hand, antenna selection, channel assignment, and power control for the cellular mode users can be centrally controlled by the cloud. The cloud enforces a mode selection and operation policy rather than allocating channels and power levels for each D2D user. The policy typically dictates the conditions at which users select their operation mode and/or enforces interference limits for D2D users on cellular users. One approach to supervise the distributed control of network functions is through bias factors, which encourage/discourage users to take certain actions, as discussed in the next subsection.

Biased Network Functionalities

The association strategy can be manipulated via tunable bias factors that artificially encourage users to associate to a certain tier for each of the uplink, downlink, and control signaling. These bias factors can expand the coverage regions of small cells to proportionally balance the load served by each network tier. Also, a control association bias factor toward higher network tiers (i.e. micro and macro), which is proportional to user mobility, can be used to reduce handover signaling.

It is worth mentioning that CoMP can be considered as a multiple BS association that can also be controlled via bias factors. In this case, each user is responsible to report the set of candidate serving BSs to the cloud, which then manages the resource allocation for that user within the selected BSs. Tuning the bias factor that encourages/discourages cooperation controls the extent to which cooperation is enabled in the network, to tune the tradeoff between the SINR performance of CoMP and the associated backhaul traffic [10].

Similar to user association, D2D communication and multi-hop relaying can also be manipulated via bias factors that encourage/discourage transmitters to select single/multi-hop D2D communication. Setting the bias factor to zero enforces direct BS communication; setting the bias factor to a high value enforces D2D communication and multi-hopping. Hence, the bias factor can be tuned according to the traffic load and population to control the number of hops, delay, and hop distance.

Application dependent bias factors can be exploited to enforce general operator policies (i.e. objective and constraints) and guide the cognitive behavior of each application. The bias factors are

calculated and adapted in the cloud, according to the traffic's spatial and temporal variation, and then dictated to the network elements.

Limitations of the Cloud Operation

As shown in Fig. 1, the cloud acts as an engine that adapts the proposed virtualized architecture to the type of applications and network conditions. A written script running at the cloud controls the network behavior. As discussed earlier, some network functions may be centrally executed in the cloud and others may be executed in a distributed manner (i.e. by BSs or devices). Note that information about devices, BSs, and network conditions are required at the cloud to dictate the guidelines for distributed network functions and to execute centralized network functions. This information is subsequently communicated to the cloud by network entities via backhaul links, e.g. fiber optic cables and wireless backhaul links.

While fiber connections may be suited for cells of medium to large size (micro and macrocells) [11], their deployment cost becomes a major problem for hundreds or thousands of small-cell BSs, as is the case in dense 5G networks. Furthermore, even when available in abundance in urban areas, fiber optic cables may not be found at the exact location where a small-cell BS exists. In contrast, wireless backhaul links are more suitable to support small-cell networks, as they are easy to plan and deploy when compared to fiber optics [12]. However, given the scarcity of the available sub-6 GHz licensed spectrum band, investment in higher frequency ranges is needed, which itself leads to a smaller coverage and needs fine tuning (i.e. strong LoS path) between the cloud and the served entities. Most importantly, a major problem in wireless backhaul design is the latency issue, as retransmissions are often required due to unsuccessful reception [13]. Additional latency is also added due to long round trip information routing to the data centers (i.e. cloud physical locations). In this case, routing delays can be reduced via multi-cloud location planning [8].

The above factors (i.e. cost, latency, coverage) often limit the operation of the cloud, and require intelligently jointly designing both the core (clouds to BSs) and RAN (BSs to users). Besides multi-cloud cooperation and coordination, synchronization and joint provisioning of resources between the backhaul links and the RAN are promising future research directions, so as to ensure the feasibility of the virtualized cognitive architecture.

Numerical Results

In this section we show the potential gain of decoupling the uplink, downlink, and control associations. While uplink and downlink decoupling improves SINR statistics, decoupling control association reduces signaling overhead [2]. Figure 3 shows the gain in terms of outage probability if uplink and downlink associations are decoupled. We consider a two-tier cellular network with channel inversion power control, in which all users maintain an average power of ρ at their serving BSs. The figure shows that if users associate in the uplink based on the uplink RSS, rather than the downlink RSS, the SINR outage probability can be reduced by 30 percent. Note that the outage probability for both the coupled and decoupled association increases with the intensity of small cells due to the increased number of interferers and the fixed received power at the test BS (i.e. ρ).

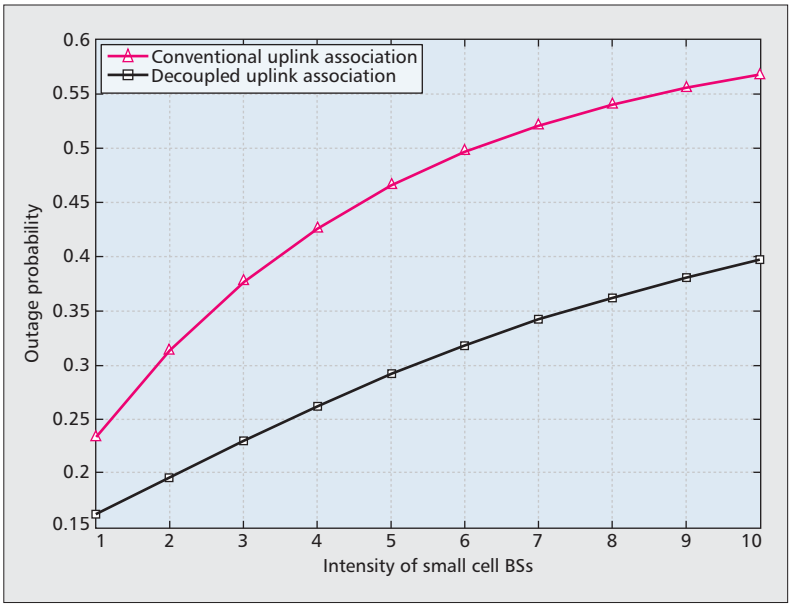


Figure 3. Outage probability vs small-cell intensity for virtualized and conventional cellular architecture, the curves are obtained via stochastic geometry analysis following the footsteps of [14].

Consider the two tier network shown in Fig. 4, in which a user follows the trajectory highlighted in black. In conventional network architecture, the test user performs a complete handover (i.e. dissociate from the serving BS, associate to the target BSs, and inform the core network for data flow switching) with every crossing over a cell boundary. Decoupling the control and data allows the macro-BS to act as a mobility anchor providing control signaling while the small cells only provide data packets (referred to as *lean carriers* in the literature) [2]. In this case, complete handovers take place in transitions between macro-cells only (i.e. the red boundary shown in Fig. 4). In contrast, only *virtual handovers* (i.e. only data packet switching between BSs) take place in transitions involving small cells. In a virtual handover, the macro cell acts as a handover anchor and the core network is not informed (i.e. the MME and SGW), which reduces the handover delay and core network signaling. Further, macro BSs have large coverage areas, which decreases the complete handover rate. In fact, the complete handover rate becomes independent of the small cell intensity. To show the amount of handover reduction by a virtualized network architecture, we plot Fig. 5. The figure shows that the complete handover rate linearly increases in the small cell intensity. The handover rate is dominated by horizontal small cell to small cell handover. On the other hand, macro to macro handover reduces as the macro boundaries are populated with small cells.

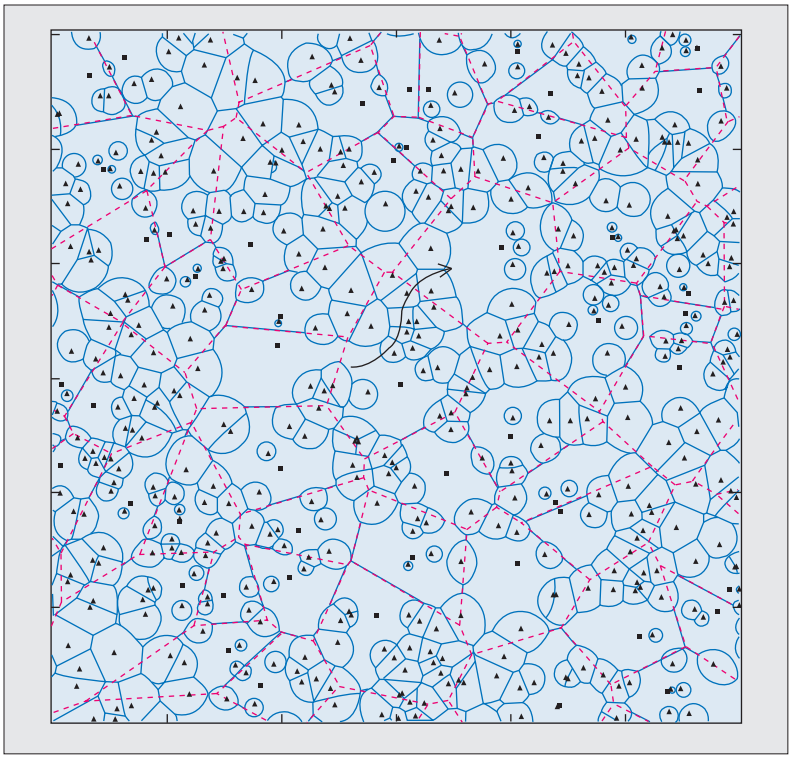


Figure 4. A two tier cellular networks with macro-BS (squares), small-cells (triangles), and a user's trajectory (highlighted in black). The figure shows the handover boundaries (in blue) for the conventional cellular network architecture and handover boundaries (in dotted red) for the virtualized cellular network architecture.

For the virtualized network architecture, the complete handover rate is constant. Hence, the visualization gain increases with the intensity of small cells. Note that reducing the handover rate can be directly translated to reduced delay and increased throughput. Hence, higher capacity gains can be harvested from network densification. It is worth noting that cognition may play an important rule in selecting the BS providing the control information, because in a dense small-cell deployment, macro-BSs may have insufficient BW to provide control signaling for all users. Hence, control signaling for high mobility users only is handled by macro cells. Control signaling for lower mobility profiles is handled by smaller BSs (i.e. micro and pico).

Conclusion

Architectural ossification for cellular networks is a limiting performance obstacle. Cognitive and flexible network operation via data/control plane decoupling and CRAN is an appealing trend to overcome the architectural ossification problem. In this case the cellular networks evolve from a deterministic infrastructure to a flexible, cognitive, and context aware architecture. Such an evolution is expected to improve network performance and make possible the foreseen 5G performance gains. To this end, this article discussed implementation of different network functions and shed light on the tradeoffs between complexity, signaling, and performance. The article proposes a distributed network control scheme in which the cloud may allow a guided distributed execution for network functions in which the guidelines are dictated by the cloud according to the application, operator policies, traffic, and network conditions. The article also showed that bias factor based guidance for network functions can be used to achieve certain design objectives. Finally, the article presented a case study to show the performance gain from decoupling uplink, downlink, and control association in multi-tier cellular networks.

Acknowledgment

The work of M.-S. Alouini was supported by the Qatar National Research Fund (a member of the Qatar Foundation) under NPRP Grant NPRP 5-250-2-087. The work of T. Y. Al-Naffouri was a supported by KAUST project no. EE002355 at the Research Institute, King Fahd University of Petroleum and Minerals. The statements made herein are solely the responsibility of the authors

References

[1] J. G. Andrews *et al.*, "What Will 5G Be?," to appear in *IEEE JSAC*, issue on 5G Wireless Commun. Systems, Sep. 2014, on archive: http://arxiv.org/pdf/1405.2957v1.pdf.

[2] Y. Kishiyama *et al.*, "Future Steps of LTE-A: Evolution Toward Integration of Local Area and Wide Area Systems," *IEEE Wireless Commun. Mag.*, vol. 20, no. 1, Feb. 2013, pp. 12–18.

[3] 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.

[4] K. Hongseok *et al.*, "Distributed–Optimal User Association and Cell Load Balancing in Wireless Networks," *IEEE/ACM Trans. Net.*, vol. 20, no. 1, Feb. 2012, pp. 177–90.

[5] H.-S. Jo *et al.*, "Heterogeneous Cellular Networks with Flexible Cell Association: A Comprehensive Downlink SINR Analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, Oct. 2012, pp. 3484–95.

[6] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Commun. Surveys and Tutorials*, 2014, pp. 1–19.

[7] H. Tabrizi, G. Farhadi, and J. M. Cioffi, "A Framework for Spatial Reuse in Dense Wireless Areas," *IEEE Global Telecommun. Conf. (Globecom 2013)*, Atlanta, GA, USA, Dec. 2013.

[8] A. Blenk *et al.*, "Applying NFV and SDN to LTE Mobile Core Gateways; The Functions Placement Problem," *ACM Special Interest Group on Data Communication (SIGCOMM'14), 4th Workshop on All Things Cellular: Operations, Applications and Challenges 2014*, Chicago, USA, Aug. 2014.

[9] C. Yu *et al.*, "On the Performance of Device-to-Device Underlay Communication with Simple Power Control," *Proc. IEEE 69th Vehic. Tech. Conf. (VTC Spring 2009)*, Barcelona, Spain, 2009.

[10] A. Sakr, H. ElSawy, and E. Hossain, "Location-Aware Coordinated Multipoint Transmission in OFDMA Network," *IEEE Int'l. Commun. Conf. (ICC 2014), Sydney*, Australia, June 2014.

[11] Y. Shi, J. Zhang, and K. B. Letaief, "Group Sparse Beamforming for Green Cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, May 2014, pp. 2809–23.

[12] H. Dahrouj *et al.*, "Coordinated Scheduling for Wireless Backhaul Networks with Soft Frequency Reuse," *European Signal Proc. Conf. (EUSIPCO)*, Sept. 2013, pp. 174–77.

[13] D. Chen, T. Quek, and M. Kountouris, "Wireless Backhaul in Small Cell Networks: Modelling and Analysis," I*EEE 79th Vehic. Tech. Conf. (VTC Spring)*, May 2014, pp. 1–6.

[14] H. ElSawy and E. Hossain, "On Stochastic Geometry Modeling of Cellular Uplink Transmission with Truncated Channel Inversion Power Control," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, Aug. 2014, pp. 4454–69.

[15] J. Bao and B. Liang, "Stochastic Geometric Analysis of User Mobility in Heterogeneous Wireless Networks," to appear in *IEEE JSAC on Recent Advances in Heterogeneous Cellular Networks*, Apr. 2015, http://www.comm.utoronto.ca/liang/research.html.

Biographies

Hesham ElSawy [S'10, M'14] (hesham.elsawy@kaust.edu.sa) received the B.Sc. degree in electrical engineering from Assiut University, Assiut, Egypt, in 2006, the M.Sc. degree in electrical engineering from the Arab Academy for Science and Technology, Cairo, Egypt, in 2009, and the Ph.D. degree in electrical engineering from the University of Manitoba, Winnipeg, MB, Canada, in 2014. Currently he is a postdoctoral fellow with the Computer, Electrical, and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Saudi Arabia, and an adjunct faculty at the school of Computer Science & Engineering, York University, Canada. From 2006 to 2010 he worked at the National Telecommunication Institute, Egypt, where he conducted professional training both at the national and international levels, as well as research on network planning. From 2010 to 2014 he worked with TRTech, Winnipeg, MB, Canada, as a student researcher. For his academic excellence, he has received several academic awards, including the NSERC Industrial Postgraduate Scholarship during the period of 2010 to 2013, and the TRTech Graduate Students Fellowship in the period of 2010 to 2014. He has also been recognized as an exemplary reviewer by the *IEEE Transactions of Communication*. His research interests include statistical modeling of wireless networks, stochastic geometry, and queueing analysis for wireless communication networks.

Hayssam Dahrouj [S'02, M'11, SM'15] (hayssam.dahrouj@gmail.com) received his B.E. degree (with high distinction) in computer and communications engineering from the American University of Beirut (AUB), Lebanon, in 2005, and his Ph.D. degree in electrical and computer engineering from the University of Toronto (UofT), Canada, in 2010. In May 2015 he joined the Department of Electrical and Computer Engineering at Effat University as an assistant professor, and also became a visiting assistant professor at King Abdullah University of Science and Technology (KAUST). Between April 2014 and May 2015 he was with the Computer, Electrical and Mathematical Sciences and Engineering group at KAUST as a research associate. Prior to joining KAUST he was an industrial postdoctoral fellow at UofT, in collaboration with BLiNQ Networks Inc., Kanata, Canada, where he worked on developing practical solutions for the design of non-line-of sight wireless backhaul networks. His contributions to the field led to five patents. During his doctoral studies at UofT he pioneered the idea of coordinated beamforming as a means of minimizing intercell interference across multiple base stations. The journal paper on this subject was ranked second in the 2013 IEEE Marconi paper awards in wireless communications. His main research interests include cloud radio access networks, cross-layer optimization, cooperative networks, convex optimization, distributed algorithms, and free-space optical communications.

Tareq Y. Al-Naffouri [M'10] (tareq.alnaffouri@kaust.edu.sa) received his B.S. degrees in mathematics and electrical engineering (with first honors) from King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 1994, his M.S. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1998, and his Ph.D. degree in electrical engineering from Stanford University, California, in 2004. He was a visiting scholar at the California Institute of Technology, Pasadena in 2005 and a Fulbright scholar at the University of Southern California in 2008. He is currently an associate professor in the Electrical Engineering Department, King Fahd University of Petroleum and Minerals and jointly at King Abdullah University of Science and Technology (KAUST). His research interests lie in the areas of sparse, adaptive and statistical signal processing and their applications in wireless communications and in multiuser information theory.

Mohamed-Slim Alouini [S'94, M'98, SM'03, F'09] (slim.alouini@kaust.edu.sa) was born in Tunis, Tunisia. He received the Ph.D. degree in electrical engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1998. He served as a faculty member at the University of Minnesota, Minneapolis, MN, USA, then at Texas A&M University at Qatar, Education City, Doha, Qatar before joining King Abdullah University of Science and Technology (KAUST), Thuwal, Makkah Province, Saudi Arabia as a professor of electrical engineering in 2009. His current research interests include the modeling, design, and performance analysis of wireless communication systems.

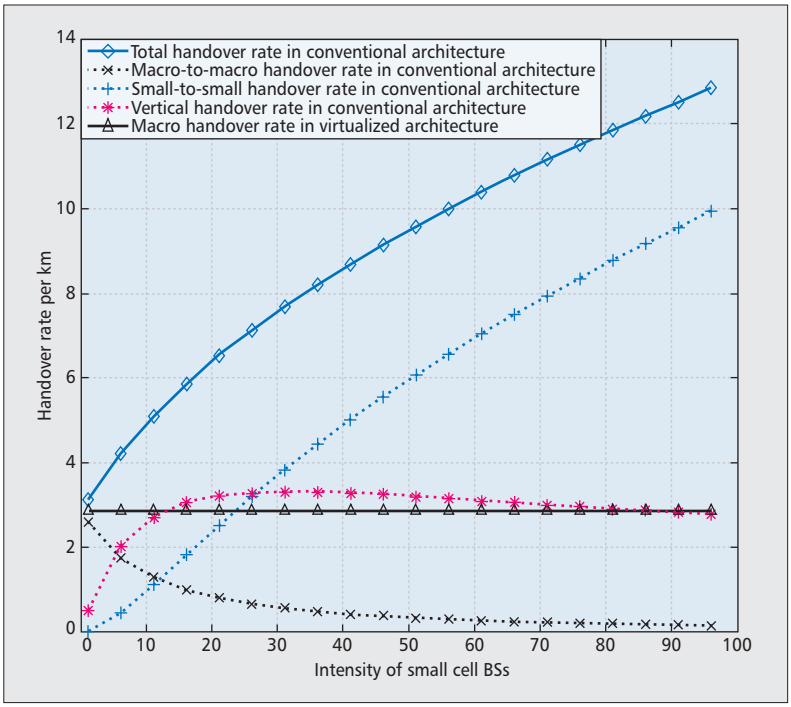


Figure 5. The handover rate per km for conventional and virtualized network architecture, the curve is plotted via stochastic geometry analysis following the footsteps of [15].

Enhanced Multi-Parameter Cognitive Architecture for Future Wireless Communications

Feifei Gao and Kaiqing Zhang

ABSTRACT

The very original concept of cognitive radio raised by Mitola targets at all the environment parameters, including those in the physical, MAC, and application layers, as well as the information extracted from reasoning. Hence, the first CR is also referred to as "full cognitive radio." However, due to its difficult implementation, the FCC and Simon Haykin separately proposed a very simplified definition, in which CR mainly detects a single parameter, spectrum occupancy, and is also called "spectrum sensing cognitive radio." With the rapid development of wireless communications, the infrastructure of a wireless system becomes much more complicated, while the functionality at every node is desired to be as intelligent as possible, say self-organized capability approaching 5G cellular networks. It is then interesting to reconsider Mitola's definition and think whether one could, besides obtaining the on/off status of the licensed user only, achieve more parameters in a cognitive way. In this article, we propose a new cognitive architecture targeting multiple parameters for future cellular networks, which is one step further toward "full cognition" compared to most existing CR research. The new architecture is elaborated in detailed stages, and three representative examples are provided based on recent research progress to illustrate the feasibility as well as the validity of the proposed architecture.

INTRODUCTION

The rapid development of wireless communications has engendered rapid proliferation of media-rich mobile devices as well as significant enhancement of communication ability. The actual mobile traffic in 2010 was five times greater than an official forecast made by the International Telecommunication Union (ITU) in 2005, and will continuously increase about 1000 times till the end of 2020 [1].

System capacity could be enhanced by enlarging network coverage (via relays, femtocells), increasing the space dimension (via massive multiple-input multiple-output [MIMO] technology), improving bandwidth efficiency and energy efficiency (via cognitive radio [CR] technology, green communications) [2]. On the other side, dynamic networks are being adopted to speed up service innovation in a more intelligent way. For example, self-organization networks (SONs) with artificial intelligence serve as a promising solution to address the challenges imposed by large-scale networks, such as the high cost of configuring and managing networks, the fluctuating nature of the available spectrum, and the diverse quality of service (QoS) requirements of various applications [3, 4]. In order to realize these key technologies, cross-layer parameters, for example, spectrum occupancy, transmit power, modulation, constellation, channel coding, location, and cell edge, should be achieved and shared among nodes in the network. However, the current information exchange scheme still relies on cooperative feedback among nodes, which causes severe transmission overhead especially when the scale of the network and the amount of data traffic are large. Hence, an intelligent way to cognitively obtain as many parameters as possible at every node could greatly enhance the communication efficiency.

The term "full cognitive radio" was already proposed by Mitola [5] in which the cognition targets every possible parameter observable by a wireless node or network. However, this ultimate goal of CR only stays at the conceptual level, while the last decade of studies in parameter achievement mainly focused on one single parameter: the occupancy of a particular spectrum. Various parameters depicting the network status — channel occupancy, transmit power level, signal constellation, modulation scheme, and channel coding, as well as cell coverage, network topology, user preferences, communication protocols, and sensing policies — have not been considered under a unified framework. It is then necessary to expand the parameter space from a single "spectrum hole" to a "multiple parameters" space in order to make the network more "intelligent."

The authors are with Tsinghua University.

0163-6804/15/$25.00 © 2015 IEEE

Meanwhile, the multi-parameter space also shares highly structured and mutually coupled characteristics, which could be further utilized to enhance cognition performance and reduce cognition complexity. If sufficient prior knowledge is available at each node, the cognition of multiple parameters will be similar to the conventional estimation and detection problem for single-parameter cognition, but with an enlarged dimension of parameter space. In this case, modified signal processing techniques (e.g., multiple hypotheses testing) could be applied [6]. Nonetheless, in complicated networks, prior knowledge is normally unavailable, and conventional signal processing techniques can hardly be executed. A solution for multiple parameters cognition is then to introduce machine learning theory, which is able to establish a prior knowledge base automatically from the proliferation of traffic data. The amount of data could be accumulated quickly considering the signal transmission speed over mega- or gigabits. In fact, machine learning has already emerged as a booming research area, the core idea of which is to mine the "patterns" behind massive data and further identify or predict unknown patterns. The inherent coupling and inter-relationship of network parameters can be recognized as patterns through learning techniques in this context.

Practically, the parameters that determine the control and optimization policies in networks change frequently, while a small change in environmental parameters may result in a big change in systematic behaviors [4]. Parameter cognition should then be as adaptive and predictive as possible to cope with the dynamic characteristics of parameters. If the cognition is able to provide accurate prediction of the network status from the patterns behind multi-parameter space, better quality of service (QoS) can be achieved via resource pre-distribution.

In this article, we propose a new cognitive architecture based on either conventional signal processing or machine learning techniques, targeting multiple parameters in future wireless communication networks. The remainder of the article is organized as follows. We first describe the core idea and the framework of the proposed cognitive architecture. We also provide three concrete examples that demonstrate the validity and efficacy of the architecture. A summary and some prospects are highlighted at the end of the article.

Architecture of Enhanced Multi-Parameter Cognitive Techniques

The architecture of the proposed cognitive technique is elaborated in three main stages: parameter structuralization, multi-parameter cognition, and parameter prediction.

Parameter Structuralization

As illustrated in Fig. 1, a communication network is described by a gigantic number of parameters at different layers, and cognition for as many parameters as possible at each node could greatly release the expense of feedback and enhance network efficiency.

Several efforts have been made in parameter cognition such as channel estimation and spectrum detection, and some fine results have been established in the past decades. However, each of these efforts concerns the extraction of only an individual parameter, ignoring the potential relationship among different parameters (i.e., parameter structure). The structure of parameters in the same layer can be achieved through direct signal processing techniques, while structure at different layers can be achieved through pattern recognition and learning techniques. For example, the cognized symbol constellation will directly exhibit the spectrum occupancy status, and could also indicate the transmission behaviors or preferences of users through a certain data mining approach.

There are three main advantages of parameter structuralization:

- Structuralization can improve the cognition accuracy since various parameters complement one another in light of the "structure."
- Complexity of cognition can be reduced since a smaller number of parameters need to be considered due to the correlation among them.
- The range of parameter cognition could be expanded because some non-cognizable parameters can be achieved by reasoning from the cognizable parameters via machine learning techniques.

Multi-Parameter Cognition

The proposed multi-parameter cognitive architecture is shown in Fig. 2, where the most popular parameters include spectrum occupancy, transmit power, modulation, and user behavior.

In general, parameter cognition can be categorized into two cases, prior-sufficient and prior-deficient, depending on how much prior knowledge is available before cognition. For the prior-sufficient case, where the key information (e.g., interference level, noise characteristics, channel statistics) is known to users, the multiple parameters cognition problem could be solved by conventional signal processing techniques. In this case, closed-form solutions normally exist, which sometimes serve as the performance bound for other low-complex suboptimal algorithms.

When the prior knowledge is insufficient to execute conventional signal processing algorithms, one may resort to learning techniques (e.g., data-driven algorithms from machine learning and pattern recognition) to intelligently establish the cognition knowledge base. Indeed, some machine-learning-based algorithms have been proposed for various tasks in CR parameter cognition. For example, in [7], a spectrum sensing engine based on a support vector machine (SVM) was designed, advancing sensing performance with smaller samples compared to the energy detector. In the cooperative spectrum sensing paradigm, [8] presents reinforced learning methods to reduce the sensing overhead by releasing network flow congestion. Most of the learning-based techniques, however, are mainly

When the prior knowledge is insufficient to execute conventional signal processing algorithms, one may resort to learning techniques, such as data-driven algorithms from machine learning and pattern recognition, to intelligently establish the cognition knowledge base.

Following the multiple parameters cognition, users are able to utilize internal structure as well as underlying pattern of parameters to further understand the parameter evolution. For example, channel occupancy status can be predicted by learning the traffic characteristics of licensed systems using the neural network model.

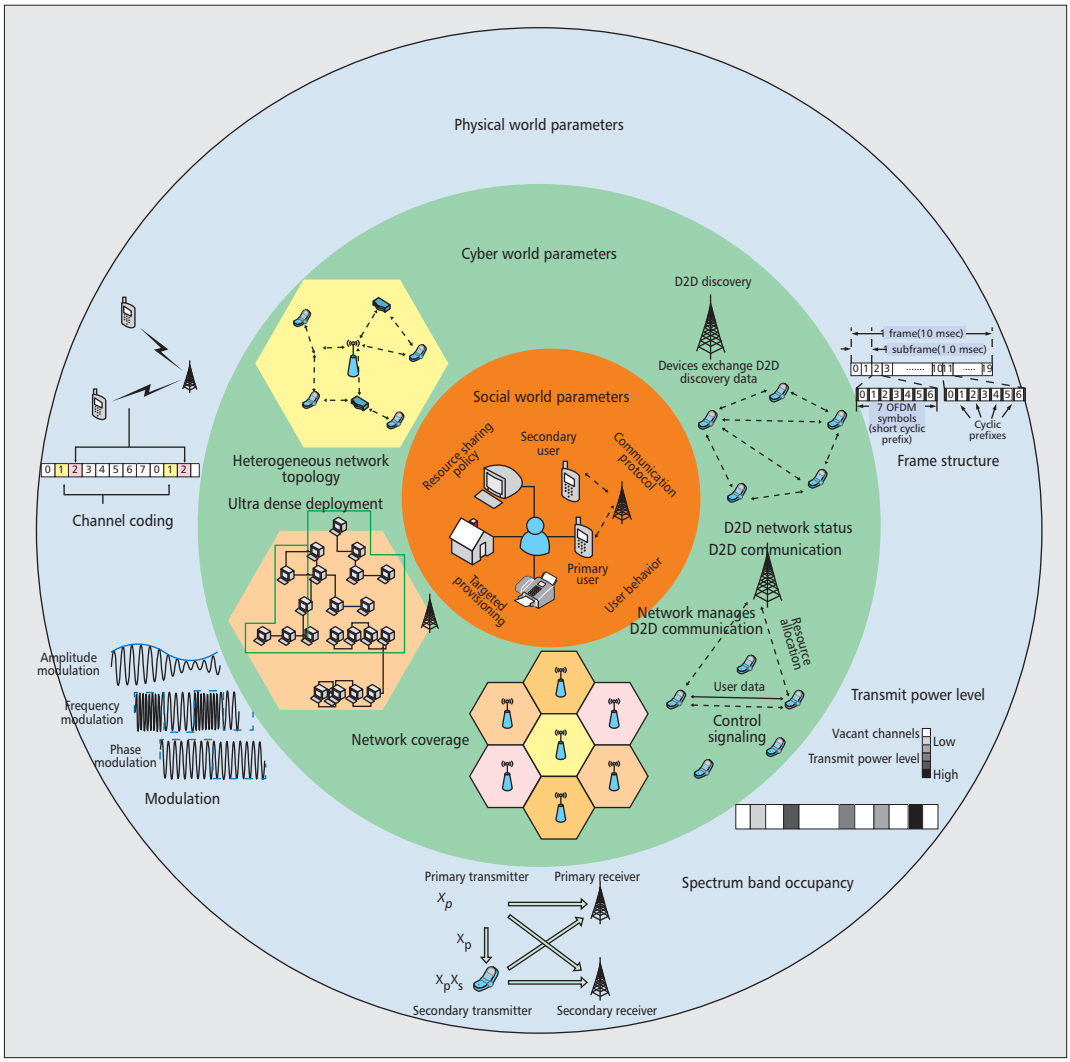


Figure 1. The enormous number of parameters involved in cognitive communication networks.

designed to obtain a single parameter (e.g., spectrum occupancy) and do not consider the learning of multiple system parameters. Moreover, these learning-based techniques fail to perform reasoning and inferring among parameters, which is known as the senior functionality of artificial intelligence, due to little attention being paid to the parameter structure.

Learning-based cognition techniques seem to be more applicable in real CR setups and are subject to little performance downgrade in the above cases. Nonetheless, there are still some deficiencies that are worth further attention because there is no "free lunch." First, it is challenging to collect sufficient "clean" data for learning and recognition. Extra overhead and high complexity for establishing and maintaining the database during the whole cognition process are unavoidable. Second, in some centralized control and optimization scenarios, learning techniques may not be adequate to cope with large-scale traffic data sets. Hence, it may be necessary to incorporate non-parametric learning algorithms with low computational complexity for cognition. Moreover, data-driven algorithms can be prone to data falsification and attack from malicious users, and more efforts are necessary for safety analysis and protection operations.

Parameter Prediction

Following multiple parameters cognition, users are able to utilize internal structure as well as underlying pattern of parameters to further understand the parameter evolution. For example, channel occupancy status can be predicted by learning the traffic characteristics of licensed systems using the neural network model [9]. The reliable prediction of channel status considerably reduces the overhead consumed by spectrum sensing because only those channels predicted to be idle need to be sensed in the next time slot. Moreover, if the unlicensed user can predict other parameters of licensed users (e.g., power level, modulation, and coding-scheme), it could then adjust the corresponding parameters such that the interference caused to the licensed user can be optimally reduced. In intelligent communication networks (e.g., SONs), predictability normally serves as the counterpart of scalability but is difficult to model due to the complexity of systems [4]. In this case, learning-based cognition is much more applicable for parameter prediction because of its extrapolative nature and simple implementation.

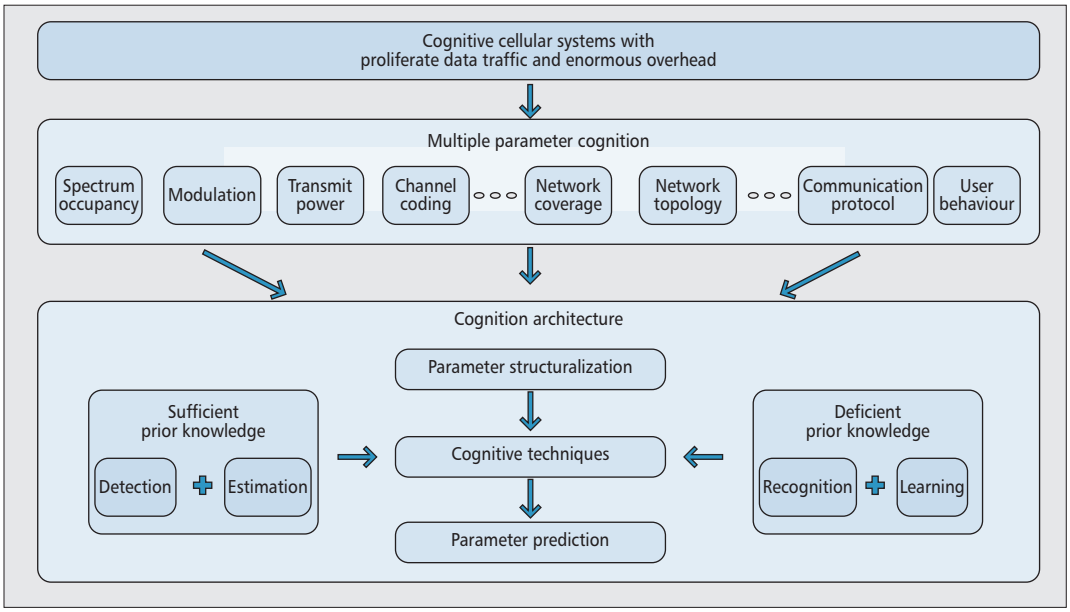


Figure 2. Framework of the multi-parameter space cognitive architecture.

EXAMPLES

In the following, we provide three representative examples of the latest research progress to show how signal processing and machine learning techniques could be used to cognize multiple parameters.

EXAMPLE 1: COGNITION FOR TRANSMIT POWER LEVELS

We first consider a practical CR scenario where the licensed user can work under more than one discrete power levels, as opposed to a single power level in conventional CR. The parameter that could be cognized is then not only the on/off status of the licensed user, but also its transmit power levels, where non-zero power levels indicate the "on" of a licensed user and vice versa.

If the candidate power levels and noise statistics are known in advance, the spectrum occupancy as well as the current transmit power levels can be obtained via multiple hypotheses detection [6]. Besides the false alarm probability and detection probability used to evaluate conventional spectrum sensing, one may further define a new performance metric called discrimination probability, which describes the capability of correctly recognizing each power pattern. A numerical result is shown in Fig. 3, where the theoretical curves can serve as an upper bound of performance for any suboptimal detectors.

However, the assumption that the unlicensed user fully knows all prior information cannot be true in a realistic setup. In this case, machine-learning based cognition techniques, such as clustering analysis and classifier construction, could be exploited to replace the multiple hypotheses testing approach. The unlicensed users can collect energy statistics and formulate energy feature vector so that licensed transmission patterns can be discovered by clustering analysis from a sufficient number of energy feature vectors. Specifically, feature vectors that share the same transmit power of the primary user (PU) can be grouped together, and the number of power states and the corresponding vector clusters are determined. Hence, the transmit power levels can be evaluated by the average of the vectors in each cluster, from which the knowledge about transmit patterns and preferences (the power states of the licensed user, the average power level of each state, etc.) are learned. Moreover, each energy feature vector would be labeled by the cluster number it is partitioned to. Classifiers and decision boundaries between different power states can then be constructed through supervised learning on the basis of these labeled data. A numerical example where SVM is used to train the classifiers and perform sensing tasks is shown in Fig. 4.

Once the power level is detected, the unlicensed users are able to adjust their own transmit power such that the interference temperature for that particular power level is fulfilled,[1] by which means the unlicensed users could fully squeeze the tolerance of the licensed user and maximize their own throughput.

Once the power level is detected, the unlicensed users are able to adjust their own transmit power such that the interference temperature for that particular power level is fulfilled, by which means the unlicensed users could fully squeeze the tolerance of the licensed user and maximize their own throughput.

EXAMPLE 2: COGNITION FOR MODULATION PATTERN

In this example, we address the problem of modulation recognition (MR) in the CR scenario. For conventional MR, there are two main genres of methods: the theoretical maximum-likelihood-based methods and the statistical pattern-recognition-based methods [11, 12].

However, MR in CR differs from the convention in several ways. First, the existing MR methods assume prior knowledge of the candidate signal constellations, which can be denoted as the modulation dictionary. It is obligatory for the dictionary to contain as many potential modulation types as possible, while a redundant dictionary would definitely degrade the recognition performance, especially in low signal-to-noise-ratio (SNR) regions. Hence, the modulation dic-

[1] The FCC has regulated different protection for different powered services in their recent reports [10].

tionary should be pruned tactfully before recognition. Second, the existing methods always assume the transmitter to be "on," which is not the case in the CR scenario where the channel occupancy status is also unknown, and its detection is coupled with MR. Third, the existing methods do not consider different transmit power levels, which can in fact be coupled with modulation types to characterize the transmission behavior of the licensed users.

We provide one numerical example in Fig. 5 to shed light on how to obtain the coupled parameters from an unsupervised learning approach. Higher order statistics (HOS) are used as the feature for recognition because they characterize the distribution shape of noisy baseband samples with low complexity [11]. We introduce a new concept, "modulation pattern," to denote the combination of modulation type and transmit power level. We then organize the estimates of cumulants of different orders and lags as a multiple cumulant vector, serving as the input feature of the machine learning algorithms. It can be proven that the feature vector is a multivariate asymptotic Gaussian approximation of cumulants [13]. In particular, the first cumulant in the vector should be the second-order cumulant in order to indicate the energy information.

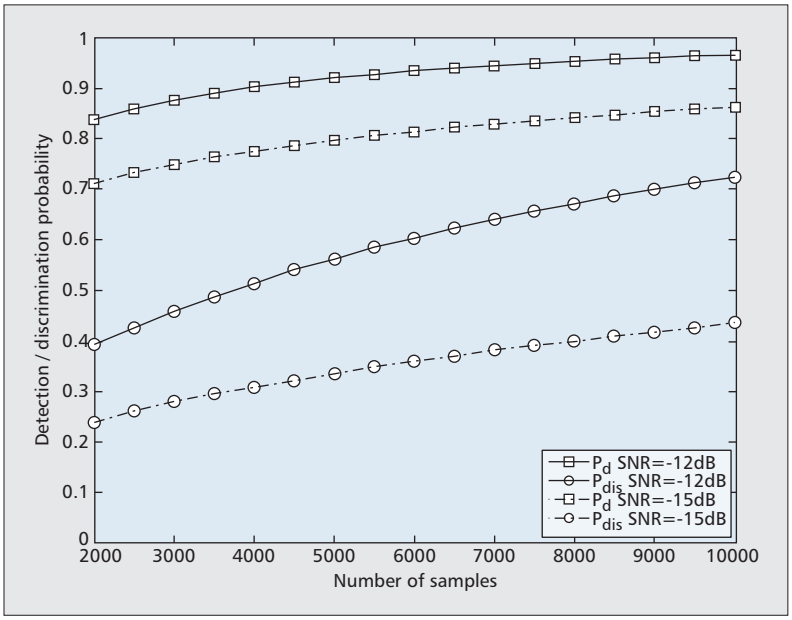


Figure 3. The theoretical performance metric of cognition vs. number of samples with system parameters known a priori.

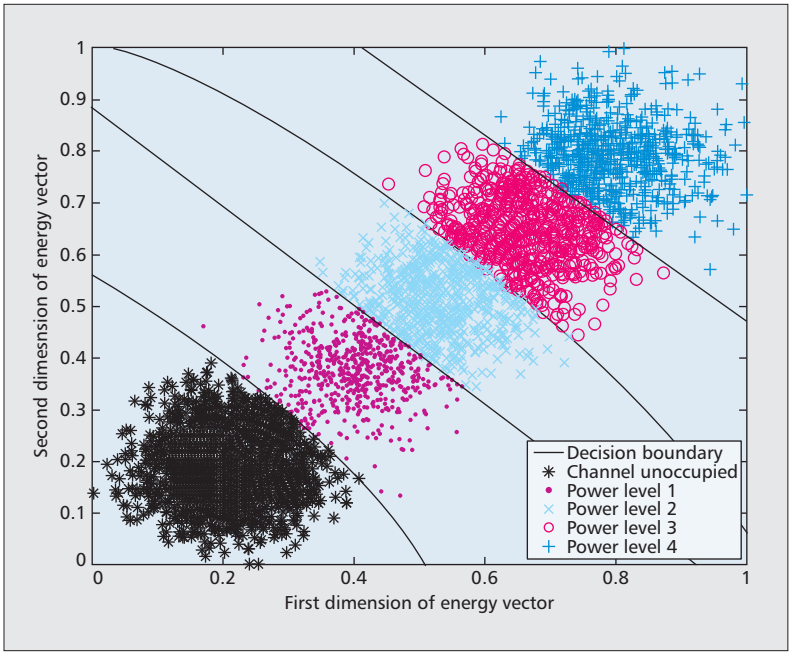


Figure 4. Clustering results and SVM trained decision boundary on normalized energy feature vectors when average SNR is –12 dB.

In order to construct the modulation dictionary and identify different "modulation patterns," we exploit the Dirichlet process Gaussian mixture model (DPGMM), a type of unsupervised clustering analysis, to construct dynamic and flexible statistical representations for the training data. We formulate a mixture model in which the number of mixture components is infinite and is not required at the beginning of clustering. After the convergence of DPGMM [14], the cumulant vectors that aggregate around the origin of coordinate system represent the "noise constellation." Detection of cumulant vectors belonging to the noise constellation reveals that only noise signals are received and no symbols are transmitted. The variance of noise can be evaluated from the second-order cumulant of the noise constellation. Furthermore, the average vector of any other cluster is evaluated. The average vectors are used to identify the alive modulation types and establish the minimal dictionary following the correspondence of HOS and modulations. Consequently, new observations of cumulant vector can be efficiently classified and predicted using the predictive probability distribution constructed by DPGMM. The dictionary can also be updated along with the update of posterior predictive distributions adaptively.

The simulation results in Fig. 5 show the efficacy of the modulation recognition and indicate superior discrimination capability compared to the pure pattern recognition approach [11]. Moreover, Fig. 5 also demonstrates that four modulation types as well as the channel idle status can be specified automatically, along which the transmit power level can be cognized as well. Furthermore, the modulation preference of a certain user is understood from each component of the predictive distribution.

Example 3: Prediction for Spectrum Environment

In the last example, we evaluate the prediction of multi-channel spectrum occupancy using machine learning techniques. Due to energy and hardware constraints, unlicensed users may not be able to perform spectrum sensing at all channels. This can be relieved by predicting the channel occupancies before each sensing time slot starts. As modeled in [15], the state of different channels and the time the licensed users spend dwelling in each state are assumed to be independent. Moreover, the change of spectrum environment can be considered as the process of

license channel vacancy and occupancy states appearing alternately. We introduce throughput as the evaluation criterion of the proposed learning-based cognitive strategy. Multi-channel spectrum sensing based on learning and prediction can be performed via three steps. First, the parameters of the learning model are estimated by the historical information embedded in the previous observations database. Subsequently, given a new observation of the channel occupancy, the next time slot can be predicted using the learning model. Meanwhile, the probability of vacancy status of each channel can be ordered from high to low. Lastly, the unlicensed user is able to efficiently perform spectrum sensing in accordance with the order and update the channel status database with the true channel status detected. The average throughput can be derived in terms of the mean error prediction probability and the probability of vacancy state for objective channels in closed form [15]. Simulation results in Fig. 6 demonstrate that multi-channel spectrum sensing and prediction based on machine learning techniques achieve better performance in terms of throughput compared to random selection of license channels.

Besides the enhancement of spectrum sensing efficiency, there are a few other benefits gained from learning and prediction of multi-channel occupancy. First, prediction results can also be utilized to foresee the traffic flow of cellular networks consisting of certain users at certain spectrum sub-bands. Hence, network congestion may be detected in advance. Moreover, prediction of channel occupancies makes it possible to predict the transmission behavior of users.

Summary and Prospects

In this article, we introduce a concrete multi-parameter cognitive architecture for future wireless communication systems that contains three key stages. Our efforts prompt a compromising but necessary way toward the ultimate goal of CR, full cognition, which currently remains at the conceptual level. Examples demonstrate that cognition in the multi-parameter space can be rather different from conventional single-parameter cognition and reveal fundamental insights on the proposed cognitive architecture. Future research could include how to utilize the cognized parameters to preserve QoS of the network in terms of data traffic, latency, overhead, and so on.

Acknowledgments

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2013CB336600, by the National Natural Science Foundation of China under Grant 61422109, by the Beijing Natural Science Foundation under Grant 4131003, and by the Tsinghua University Initiative Scientific Research Program under Grant 20121088074.

References

[1] S. Singh and P. Singh, "Key Concepts and Network Architecture for 5G Mobile Technology," *Int'l. J. Scientific Research Eng. & Tech.*, vol. 1, no. 5, Aug. 2012, pp. 165–70.

[2] C.-X. Wang *et al.*, "Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 12–30.

[3] Z. Zhang, K. Long, and J. Wang, "Self-Organization Paradigms and Optimization Approaches for Cognitive Radio Technologies: A Survey," *IEEE Wireless Commun.*, vol. 20, no. 2, Apr. 2013, pp. 36–42.

[4] Z. Zhang, K. Long, and J. Wang, "On Swarm Intelligence Inspired Self-Organized Networking: Its Bionic Mechanisms, Designing Principles and Optimization Approaches," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, 1st qtr., 2014.

[5] J. Mitola and J. Maguire, "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Commun.*, vol. 6, no. 4, Aug. 1999, pp. 13–18.

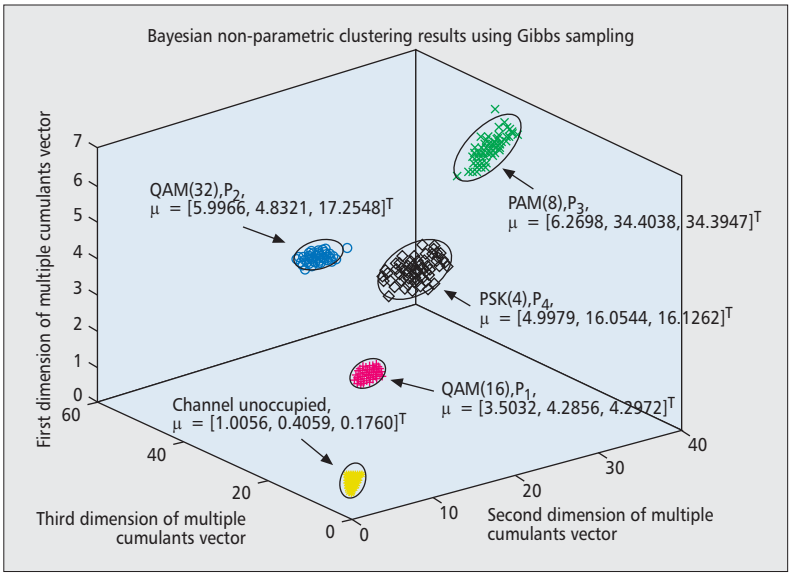


Figure 5. Pattern discovery and clustering results of multiple cumulants vectors before normlization using the DPGMM. μ denotes the mean vector of each Gaussian component, and P_i denotes the transmit power level where $P_1^2 : P_2^2 : P_3^2 : P_4^2 = 2.5 : 5 : 5.3 : 4$ when overall average SNR is 10 dB and the number of samples is 100.

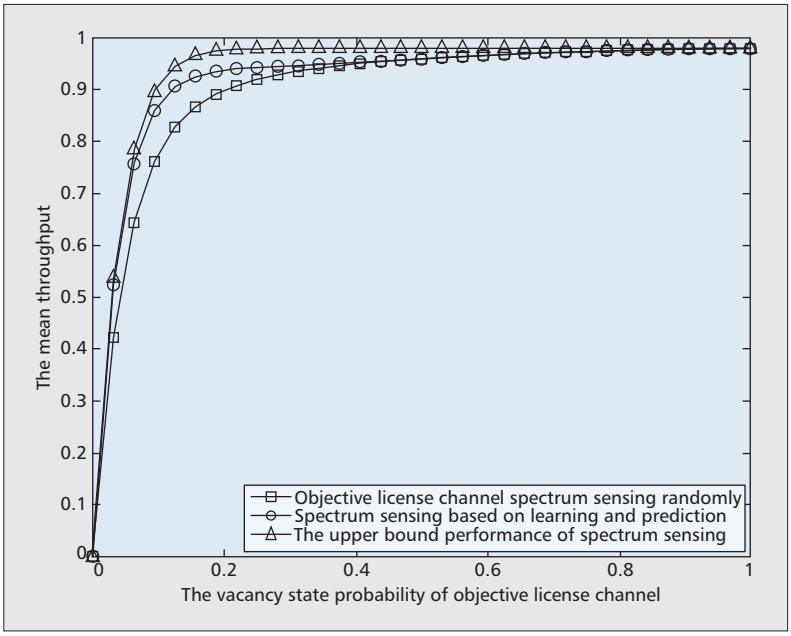


Figure 6. The performance of mean throughput vs. the probability of vacancy state when the number of channels is 25 and the normalized channel capacity is 1 b/s.

Examples demonstrate that the cognition in multiple parameters space can be rather different from conventional single parameter cognition and reveal fundamental insights in the proposed cognitive architecture. Future research could include how to utilize the cognized parameters to preserve QoS.

[6] F. Gao *et al.*, "Sensing and Recognition When Primary User has Multiple Transmit Power Levels," *IEEE Trans. Signal Processing*, vol. 63, no. 10, May 2015, pp. 2704–17.
[7] Y. Huang, H. Jiang, and H. Hu, "Design of Learning Engine based on Support Vector Machine in Cognitive Radio," *IEEE Int'l. Conf. Computational Intelligence and Software Engineering*, Wuhan, China, Dec. 2009, pp. 1–4.
[8] B. Lo and I. Akyildiz, "Reinforcement Learning-Based Cooperative Sensing in Cognitive Radio Ad Hoc Networks," *IEEE Int'l. Symp. Personal Indoor and Mobile Radio Commun.*, Instanbul, Turkey, Sept. 2010, pp. 2244–49.
[9] V. Tumuluru, P. Wang, and D. Niyato, "A Neural Network Based Spectrum Prediction Scheme for Cognitive Radio," *IEEE ICC*, Capetown, South Africa, May 2010, pp. 1–5.
[10] FCC, "In the Matter of Unlicensed Operation in the TV Broadcast Bands," ET Docket No. 04-186, Notice of Proposed Rulemaking, OET, May 2004.
[11] A. Swami and B. Sadler, "Hierarchical Digital Modulation Classification Using Cumulants," *IEEE Trans. Commun.*, vol. 48, no. 3, Mar. 2000, pp. 416–29.
[12] O. A. Dobre *et al.*, "Survey of Automatic Modulation Classification Techniques: Classical Approaches and New Trends," *IET Commun.*, vol. 1, no. 2, Apr. 2007, pp. 137–56.
[13] A. Dandawate and G. Giannakis, "Asymptotic Theory of Mixed Time Averages and *k*th-Order Cyclic-Moment and Cumulant Statistics," *IEEE Trans. Info. Theory*, vol. 41, no. 1, Jan. 1995, pp. 216–32.
[14] G. Rasmussen, "Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution," *J. Comp. Science and Tech.*, vol. 25, no. 4, July 2010, pp. 653–64.
[15] Z. Ye, Q. Feng, and K. Shen, "Spectrum Environment Learning and Prediction in Cognitive Radio," *IEEE Int'l. Conf. Signal Proc., Commun. and Computing*, Xi'an, China, Sept. 2011, pp. 1–6.

Biographies

Feifei Gao [SM] (feifeigao@ieee.org) received his B.Eng. degree from Xi'an Jiaotong University, China, in 2002, his M.Sc. degree from McMaster University, Hamilton, Ontario, Canada, in 2004, and his Ph.D. degree from National University of Singapore in 2007. He was a research fellow with the Institute for Infocomm Research, A*STAR, Singapore, in 2008, and an assistant professor with the School of Engineering and Science, Jacobs University, Bremen, Germany, from 2009 to 2010. In 2011, he joined the Department of Automation, Tsinghua University, Beijing, China, where he is currently an associate professor. His research areas include communication theory, signal processing for communications, array signal processing, and convex optimizations, with particular interests in MIMO techniques, multi-carrier communications, cooperative communication, and cognitive radio networks. He has authored/coauthored more than 60 refereed IEEE journal papers and more than 80 IEEE conference proceeding papers, which have been cited more than 2500 times from Google Scholar. He has served as an Editor of *IEEE Transactions on Wireless Communications*, *IEEE Wireless Communications Letters*, *International Journal on Antennas and Propagations*, and *China Communications*. He has also served as Symposium Co-Chair for IEEE ICC '15, IEEE GLOBECOM '14, VTC-Fall '14, as well as a Technical Committee member for many other IEEE conferences.

Kaiqing Zhang (zkq11@mails.tsinghua.edu.cn) received his B.E. degree from Tsinghua University in June 2015. He will pursue his Ph.D. degree in the Department of Electrical & Computer Engineering at the University of Illinois at Urbana-Champaign. His research interests include machine learning and optimization in networked systems with applications in smart grid.

CPC-Based Backward-Compatible Network Access for LTE Cognitive Radio Cellular Networks

Ling Liu, Yiqing Zhou, Lin Tian, and Jinglin Shi

ABSTRACT

Network access plays an important role in LTE cognitive radio (LTE-CR) cellular networks in determining users' experiences. An overview is first carried out on network access schemes in existing cognitive cellular networks such as IEEE 802.22 and IEEE 1900.4, based on which it can be seen that cognitive pilot channel (CPC)-based network access is a promising scheme for LTE-CR, which may provide fast network access and put no stringent requirements on terminals. Next, considering the implementation issues of CPC in practical systems, a CPC-based backward-compatible network access scheme should be designed for LTE-CR to facilitate the application of CR in LTE networks, which could exploit existing LTE structures and technologies to carry CPC information. To achieve this, a new system information block (SIB) should be designed to carry CPC on a current physical downlink shared channel with little standards effort. Moreover, load awareness is introduced so that LTE-CR is activated only when the system load is high. The complete process of this SIB-CPC-based backward-compatible and load-aware network access is described, and its performance is evaluated via simulations. It is shown that the blocking ratio of LTE-CR can be reduced notably compared to that of conventional LTE without CR. Moreover, by selecting an appropriate load threshold to activate LTE-CR, the average user throughput of LTE can be improved with near-zero blocking ratio by offloading users to complementary cognitive spectra.

INTRODUCTION

With the increasing demand for mobile Internet services, mobile data traffic is predicted to be increased by more than 1000 times in the next 10 years, which imposes a big challenge on mobile communication networks. As an effective technique to fully utilize the allocated spectrum resource and improve network capacity, cognitive radio (CR) [1, 2] has drawn a lot of attention in cellular networks. For example, CR is employed in IEEE 802.22 cellular systems to provide broadband wireless access using television white space (TVWS) [3]. Another application of CR is for dynamic spectrum access (DSA) in multiple-radio-access-technology (multi-RAT) heterogeneous cellular networks, which has been standardized in IEEE 1900.4 [4] and included in the end-to-end efficiency (E3) project of the European Telecommunications Standards Institute (ETSI) [5].

Recently, CR is under consideration for fourth generation (4G) mobile communication systems to complement more spectra for traffic growth. Long Term Evolution (LTE) over unlicensed spectrum such as 5 GHz and TVWS has been studied in Third Generation Partnership Project (3GPP) standardization [6–8] and is known as LTE-CR. One possible application scenario for LTE-CR is that when dual connectivity is supported, cognitive spectrum can be used for data transmission in small cells while licensed spectrum is used for the connection to macrocells. Alternatively, cognitive spectrum can also be exploited in standalone mode where the carrier supports both data transmission and control management as well as other licensed LTE carriers. In either mode, the LTE-CR network should be able to work on both licensed and harvested cognitive spectra.

However, since the LTE network becomes a secondary system in the cognitive bands belonging to other licensed networks, various challenges occur when extending LTE to LTE-CR due to the uncertainty of cognitive spectra, such as spectrum sensing, network access, and spectrum handoff. Note that network access is especially important for an LTE-CR cellular network when it works in the standalone mode. In this case, users could access the network using both the cognitive and licensed spectra, so the network access procedure is quite different from the conventional one with only licensed spectra. Some potential network access schemes have been considered for LTE-CR. One possible solu-

Ling Liu, Yiqing Zhou (corresponding author), Lin Tian, and Jinglin Shi are with the Beijing Key Laboratory of Mobile Computing and Pervasive Devices, Institute of Computing Technology, Chinese Academy of Sciences.

Ling Liu is also with the University of the Chinese Academy of Sciences.

0163-6804/15/$25.00 © 2015 IEEE

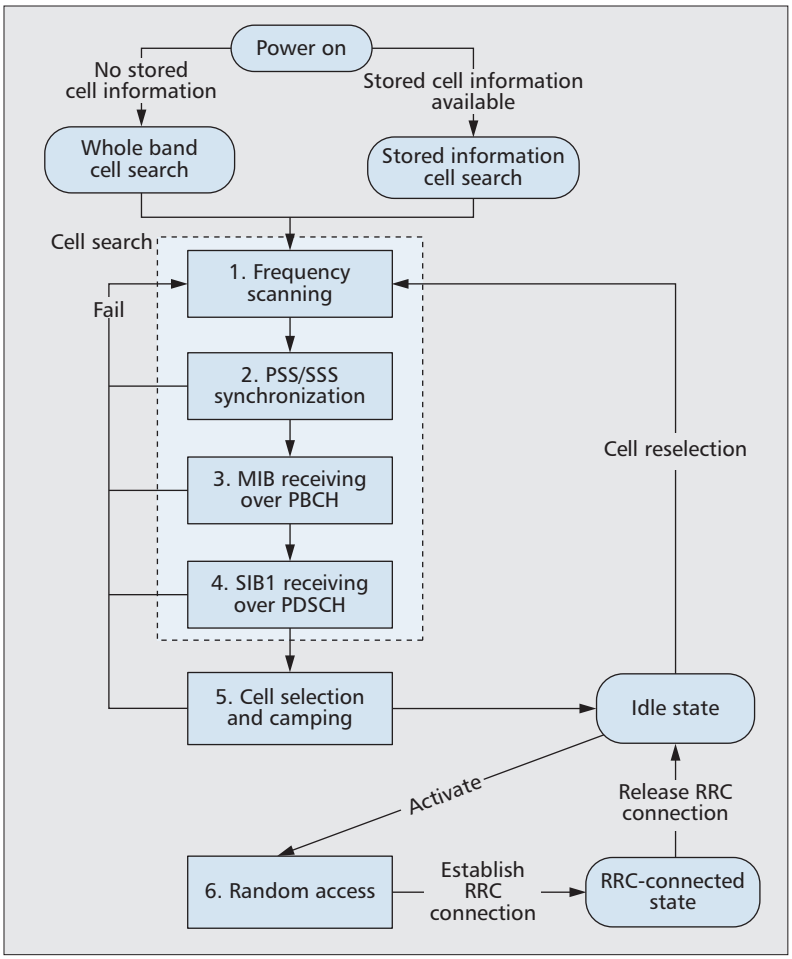


Figure 1. Network access in LTE networks.

tion is to modify the current LTE air interface so that mobile users can sense the spectrum and use existing schemes such as listen-before-talk and request-to-send/clear-to-send to access the network [7]. However, changing the LTE air interface is not easy, so it is more desirable to use the existing LTE air-interface for either licensed or unlicensed spectrum [6, 8] to reduce the operating cost and system complexity. Therefore, this article focuses on backward-compatible network access schemes for LTE-CR.

Network Access in LTE and CR Cellular Systems

Network Access in LTE

The network access process in LTE is shown in Fig. 1. Among all the procedures, the most time-consuming is whole band cell search, where no previous cell information is stored and users need to scan all the carrier frequencies for cell search. The searching step is usually 100 kHz, and the searching range could be around 1000 MHz for time-division LTE (TD-LTE) and around 862MHz for downlink of frequency division duplex LTE (FDD-LTE) [9]. At each frequency, users perform frequency scanning, synchronization, and system information (SI, including master information block [MIB] and system information block 1 [SIB1]) receiving sequentially. If the strongest signal on a frequency detected by the user is below a given threshold, there is no cell on this frequency and the user switches to the next one. On the other hand, if the signal strength is higher than the threshold, the user continues to search the primary synchronized signal (PSS) and secondary synchronized signal (SSS) for synchronization. The average time required for synchronization is several tens of milliseconds [10]. After that, the user receives a MIB over the physical broadcast channel (PBCH) for cell configuration information, which is needed to find the physical downlink shared channel (PDSCH). Next, if the user successfully receives SIB1 over PDSCH for cell selection information, cell search succeeds. Otherwise, if any step in the cell search fails, the user has to switch to the next frequency and restart the cell search.

To shorten the time of cell search, the stored information cell search is proposed, where the user firstly searches for cells based on the stored carrier frequencies used previously. On average, this could reduce the searching time significantly because users do not move frequently in practice. If the stored information cell search fails, the whole band cell search is performed.

After cell search, the user performs cell selection according to certain criteria considering both the quality of downlink and uplink received signals [9] and the cell selection information from SIB1. If the criteria are satisfied, the user camps on this cell and enters idle state. Otherwise, cell selection fails, and the user should switch to the next frequency to search for another cell.

If a user in idle state is activated, it establishes radio resource control (RRC) connection via random access and enters RRC-connected state. Thus, it can communicate with the base station (BS). Once the session is finished, the user releases RRC connection and goes back to idle state.

Network Access in CR Systems

It can be seen that in the network access of current LTE, cell search is based on the known licensed spectrum information. However, in CR systems, frequencies in the stored information may belong to the cognitive spectrum and could be occupied by primary users (PUs) with a high probability once an LTE user powers on. Thus, it may be unavailable and the time-consuming whole band cell search should be performed. Moreover, it takes a long time to detect a channel in cognitive spectra, which could deteriorate user experiences seriously. Hence, new network access techniques should be developed for LTE-CR networks to avoid such a time- and energy-consuming start-up process.

Various effective network access schemes have been proposed for practical CR cellular systems (e.g., IEEE 802.22 and IEEE 1900.4 systems) and can be classified into two categories: sensing-based (without pilot) and pilot-based schemes. Compared to pilot-based schemes, sensing-based schemes are more spectrally efficient. However, users have to sense the whole spectrum to obtain the information to access the network. Hence, it is only suitable for systems working in a limited spectrum range. For the

pilot-based scheme, users can listen to the pilot channel to obtain the available network information. Without the need to sense the spectrum, the pilot-based network access is fast and energy-efficient. But the cost is reduced spectral efficiency due to additional resource allocated to the pilot channel.

For instance, IEEE 802.22 systems, which aim to using TVWS between 54 and 862 MHz, employ sensing-based network access without pilot. This is because the number of TV channels is usually limited between 50 to 70, each with a bandwidth of 6–8 MHz [3], and the spectrum sensing time is acceptable. Moreover, most IEEE 802.22 terminals are installed in houses, so their power consumption is not a problem. However, for IEEE 1900.4 systems focusing on DSA in multi-RAT networks, the spectrum range is extremely large (e.g., from 450 MHz to 3 GHz). Thus, the time- and energy-consuming sensing-based network access is not preferred and a pilot-based network access is proposed, where a radio enabler (RE) is designed as a communication channel to carry the network information such as the available spectrum information, the load situation of the network, the objective of radio resource allocation optimization, and so on [4]. After detecting RE, a user could decide which frequency and which RAT to access according to its requirement and the guidance of the resource optimization policy.

Similar to the RE in IEEE1900.4, a cognitive pilot channel (CPC) is employed in the E3 project that provides access information like the available frequencies, relevant load, and access strategy of each RAT. Thus, a user can access a cell by listening to the CPC after it powers on. The implementation of CPC is also under standardization of the International Telecommunication Union Radio Standards Sector (ITU-R) to facilitate the application of CR in land mobile services [11], including WiMAX and LTE.

In LTE-CR networks, cognitive spectra may include TVWS; industrial, scientific, medical (ISM) spectra (902–908 MHz, 2.4–2.5 GHz and 5.725–5.875 GHz), unlicensed national information infrastructure (UNII) spectra (5.15–5.35 GHz and 5.47–5.825 GHz), and mobile communication spectra used by various RATs such as Global System for Mobile Communications (GSM), wideband code-division multiple access (WCDMA), and LTE. Thus, the spectrum range is large and the sensing time could be high. Moreover, mobile terminals feature simple hardware design, limited cognitive ability, and limited battery life. Thus, for mobile terminals, it is desirable to avoid energy-consuming spectrum sensing during network access. As a whole, pilot-based schemes are more suitable for LTE-CR networks, which provide fast network access with low power consumption and put no stringent requirements on mobile terminals. This paper focuses on the CPC based network access for LTE-CR networks.

Cognitive Pilot Channel

Basic Configurations of CPC

Although the basic idea of the CPC is easy to follow, various challenges occur when implementing CPC.

Spectrum resource for CPC: In multi-RAT networks, for the convenience of searching, it is desirable that the CPC should be global and common to all users. In this case, the CPC is a new and standalone channel and does not belong to any RAT, which is known as out-of-band CPC. However, due to the spectrum shortage, it is extremely difficult to get a global frequency for out-of-band CPC. Hence, in-band CPC is proposed which is located in the existing bands and each RAT has its own in-band CPC. It is usually a logical channel in each RAT which can be mapped to different resources from time to time. But a lot of standardization efforts are needed for in-band CPC since the existing RATs should be modified to support the CPC.

RAT for CPC: If out-of-band CPC is considered, a new RAT could be designed for the CPC to provide user synchronization and information delivery control. For in-band CPC, the delivery of CPC can use the same RAT and reuse the infrastructure of the located RAT. However, a centralized CPC manager is needed to coordinate the CPCs in different RATs globally.

CPC information organization: In multi-RAT networks, since the covering networks and available spectra change with geometrical locations, it is necessary to organize the CPC information according to location. One method is to divide the coverage area of CPC into small meshes, each with different network information, as shown in Fig. 2. When a user powers on, it first obtains its location with positioning techniques such as global positioning system (GPS) and then receives the corresponding network information of its located mesh from CPC. To improve the accuracy of conveying CPC information, dynamic mesh division should be considered in a multi-RAT overlapped scenario [12].

Another approach is to organize the CPC information according to the coverage area of each RAT. The CPC information includes the RAT type, coverage extension, coverage area, and frequency list. If the coverage extension of a RAT is global, the network information is useful in the entire coverage area, and the coverage area is omitted. Otherwise, the network information is locally valid, and the geographic coordinates of the local coverage area is given in the coverage area item.

Transmission modes of CPC: The CPC transmission can be supported in two modes, broadcast and on-demand. For the broadcast mode, only a dedicated downlink CPC is needed, conveying information to all users. The implementation of this mode is simple, but it results in large delay and low transmission efficiency. For example, considering mesh-based CPC information organization, the network information of each mesh is broadcasted sequentially and periodically on the CPC. The period is related to various parameters such as mesh numbers, CPC information size, and CPC bandwidth [13]. The average time to receive the corresponding CPC information for a user is quite long, especially when the mesh number is large. On the other hand, for the coverage-area-based CPC information organization approach in which no mesh concept is applied, the delay of receiving CPC is not serious. However, it should be noted that in

So for mobile terminals, it is desirable to avoid the energy-consuming spectrum sensing during network access. As a whole, pilot-based schemes are more suitable for LTE-CR networks, which provide fast network access with low power consumption and put no stringent requirements on mobile terminals.

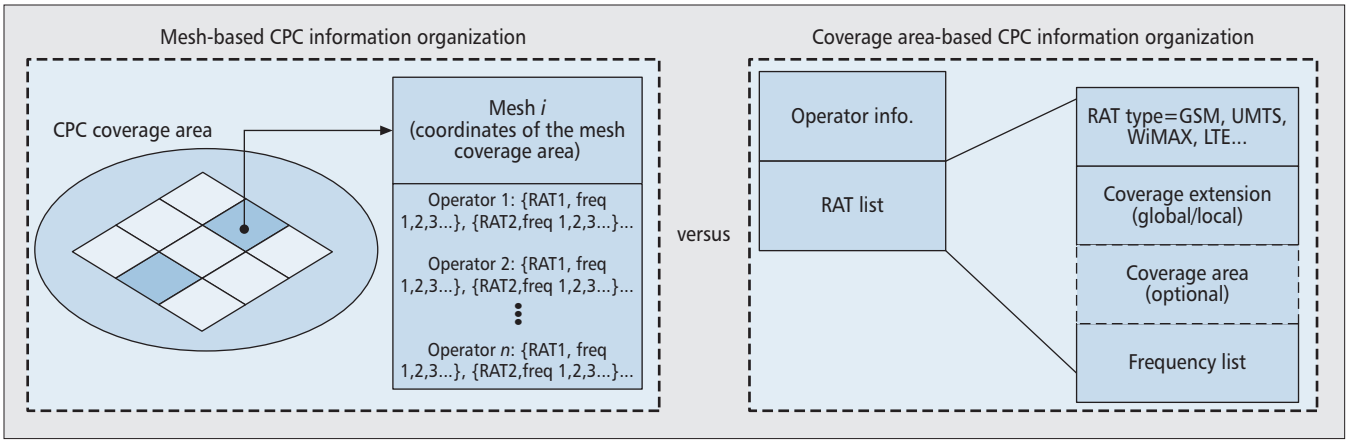


Figure 2. Two CPC information organization schemes.

broadcast mode, the system keeps broadcasting the CPC information all the time, but the information is used by users only occasionally. Hence, the transmitting efficiency is quite low.

Different from the broadcast mode, in the on-demand mode, the CPC information is transmitted to a user only when requested. Thus, the user can quickly obtain the CPC information without waiting. But the system and communication protocol become more complex in this case, where an uplink channel is needed for users to send CPC requests. Moreover, the delay of receiving CPC in on-demand mode changes with the user request rate. If the request rate is low, the delay of the on-demand CPC is much lower than that of the broadcast one. However, if the request rate gets higher, the queuing delay of CPC delivery at the network side increases and even results in a longer delay of receiving CPC than that of the broadcast one [12]. One obvious advantage of on-demand mode is the high transmission efficiency because spectrum resource is only needed when there is a request.

Key Challenges for CPC-Based LTE-CR

Based on existing research and standardization, the basic configurations of CPC for LTE-CR can be determined. First of all, the main purpose of LTE-CR networks is to exploit cognitive spectra from other networks for LTE users to improve their performance, such as lower blocking ratio and higher data rate. Thus, there is only one working RAT (i.e., LTE) for LTE-CR networks with cognitive capability. In this case, in-band CPC is more suitable, and the existing LTE technology and infrastructures can also be reused for CPC. Moreover, since there is no need to carry out mesh division for a one-RAT network, coverage-area-based CPC information organization should be adopted in LTE-CR. In addition, it is well known that in cellular networks, the bandwidth requirement in downlink is much higher than that in uplink. Hence, on-demand mode is recommended for LTE-CR, which could improve the downlink CPC transmission efficiency at the cost of an uplink request.

Given the previous basic configurations, the CPC-based network access in LTE-CR can be implemented. One possible way is to define the in-band CPC as a new channel independent of the current LTE standard. In this case, when a user powers on, a CPC search over all the licensed bands is needed. Note that the conventional signal strength detection strategy in cell search is not suitable since these channels may be used by BSs for data transmission. Therefore, specific searching strategies should be designed such as matched filtering detection with prior information of CPC signals [14]. After a CPC is found, synchronizations with CPC should be performed to receive CPC information correctly. If the existing LTE time and frequency synchronization scheme with PSS and SSS detection is reused for CPC, this may mislead users without CPC ability to synchronize with the CPC channel and affect their non-CPC synchronization. Hence, new synchronized signals should be designed for CPC. When a user obtains the network information, it should send a CPC request in uplink since on-demand CPC is employed to provide better transmission efficiency. However, at this time, the user has not established connections to the BS. Thus, an interaction scheme between the BS and the user should be developed for the initial request and response of CPC information.

According to previous analysis, it can be seen that if the CPC is defined as a new channel independent to the LTE standard, a lot of effort is needed to modify the LTE protocols to support the CPC-based network access in LTE-CR. Therefore, it is more desirable to implement the CPC-based network access based on the current LTE standard in a backward compatible way.

CPC-Based Backward-Compatible Network Access for LTE-CR

SIB-CPC-based Backward-Compatible Network Access

As discussed before, a backward-compatible network access scheme is desirable for LTE-CR that exploits existing technologies and structures of LTE to carry CPC information. It can be realized by carrying CPC in the existing physical channel PDSCH, where users receive SIB1 to carry out cell selection during conventional network access. Moreover, current LTE protocols

support the adding of a new SIB in the SI message for new functions or services. Therefore, to carry CPC information in PDSCH, a new SIB (e.g., SIB17) can be designed. After a user powers on, it first carries out cell search like that in conventional network access to synchronize to a cell operating in licensed spectra and receive SI. If the user has no cognitive capability, it continues the network access in licensed spectra according to the current protocols. On the other hand, if the user has cognitive capability, it listens to the PDSCH to receive CPC information in SIB17 and accesses to a cognitive channel. The following issues should be considered carefully when implementing the SIB-CPC-based network access.

Camping spectrum: When a cognitive channel is available, the LTE-CR user may camp on the channel in idle state. This could lead to frequent spectrum handoffs in idle state due to the uncertainty of cognitive spectra. Note that LTE networks can contain a large quantity of users in idle state (e.g., 1200) and only a limited number of users in connected state (e.g., 200 or 400) [15]. Thus, the capacity of LTE is mainly occupied by connected users that demand additional spectra to transmit data. Therefore, there should be no problem for LTE to support many idle users, and it is proposed that LTE-CR users in idle state should always camp on the licensed spectrum but not cognitive spectra to avoid frequent handoffs.

Load-aware CR access: Obviously, the communication quality on cognitive spectra is not as reliable as that on licensed spectra. In addition, LTE-CR is more complicated than a normal LTE network with additional operations including spectrum sensing, spectrum handoff, and so on. Therefore, cognitive spectra should be used only if necessary. A load-aware CR access is then proposed where access to cognitive spectra is enabled only when the cell load η, which is defined as the ratio of active user number to the maximum containable user number on licensed spectrum, exceeds a predefined threshold η_0. After the bandwidth-demanding transmission is completed, users should be back on the licensed spectra for the idle state.

CPC broadcasting: As illustrated before, access to the cognitive spectra is not always necessary. Therefore, the broadcasting of the SIB17 carrying CPC is also conditional. When the cell load is light, CR access is not enabled, so SIB17 is not needed. Once the licensed spectrum gets crowded, the BS should inform the users to receive SIB17. This can be realized by modifying the current SIB1, which is periodically broadcast and in charge of scheduling other SIBs. When the load is high and CR is enabled, users to be activated should try to find a proper cognitive channel to access. If there are no such channels, they are blocked.

Complete Scheme Description

The detailed SIB-CPC-based backward-compatible network access is shown in Fig. 3. After being powered on, a user accesses the licensed spectrum according to the current LTE cell search and cell selection mechanisms, shown in Fig. 1. Then it camps on the licensed spectra and enters idle state, keeping monitoring-relevant SI. Each BS should keep detecting the load of its cell and sensing the cognitive spectra. If the load η is light ($\eta < \eta_0$), CR is not enabled and no CPC is transmitted. In this case, if a user wants to initiate a session, it performs random access over the camped licensed spectra regardless of its CR capability. However, if the load is heavy ($\eta \geq \eta_0$), access to cognitive spectra is allowed. The BS modifies SIB1 to notify users that CPC information is carried by SIB17. In this case, if a user without CR capability is activated, it is blocked due to the lack of licensed spectra, while those with CR capability could select a cognitive channel to access. After the session on cognitive spectra is over, the user releases the connection and returns to the licensed spectrum.

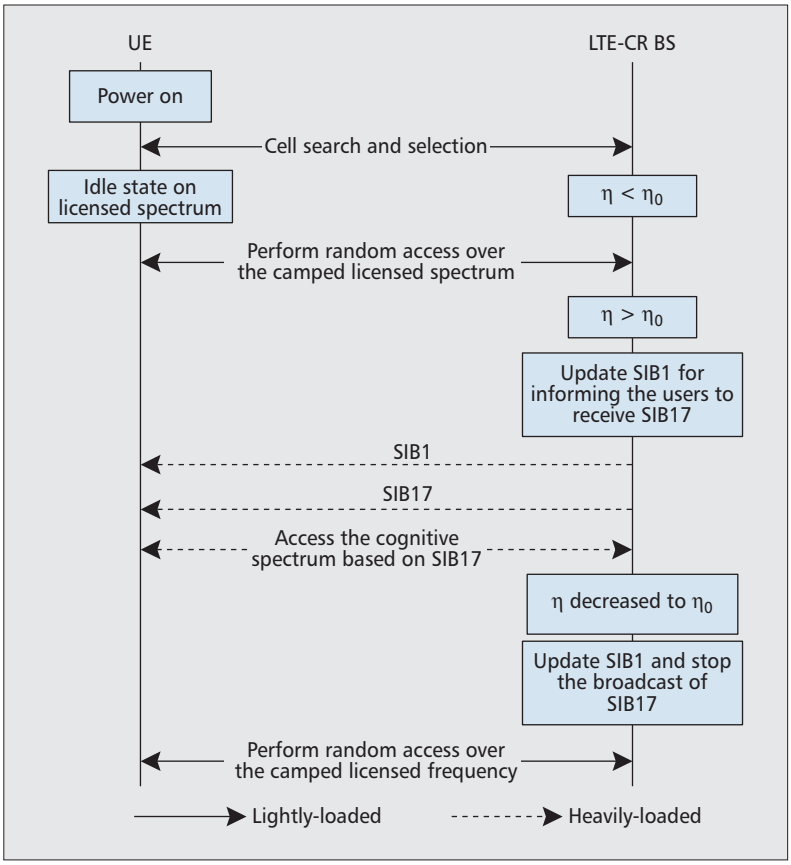


Figure 3. SIB-CPC-based backward-compatible network access.

Performance Evaluation

Simulations are carried out to show the performance of the SIB-CPC-based backward-compatible network access for LTE-CR networks. The bandwidth of LTE licensed band is 10 MHz, and TVWS is considered for cognitive access in LTE-CR. Each TV channel has a bandwidth of 8 MHz, with the central 5 MHz band for LTE-CR and the rest reserved to avoid adjacent channel interference. Assume that a 10 MHz LTE licensed channel and 5 MHz cognitive channel can support at most 20 and 10 active users, respectively. In each TV channel, the arrival and departure of PUs can be modeled as two independent Poisson random processes with mean

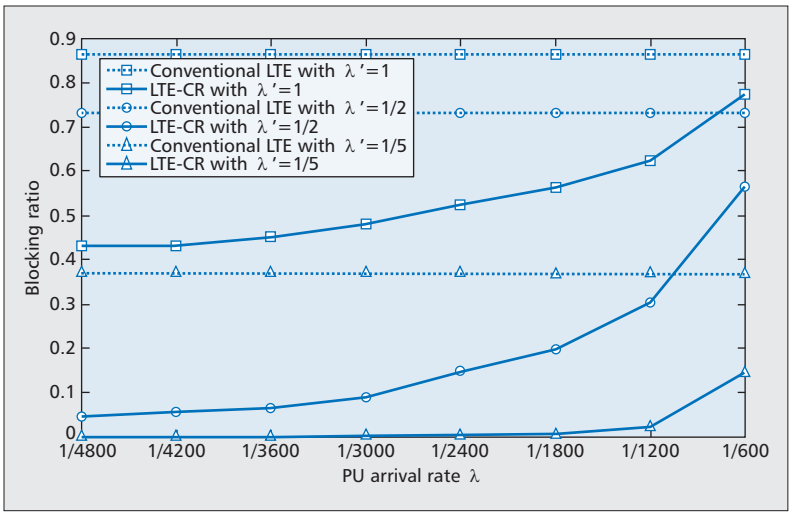


Figure 4. Blocking ratio with 10 TV channels ($\mu = 1/1200$, $\mu' = 1/150$).

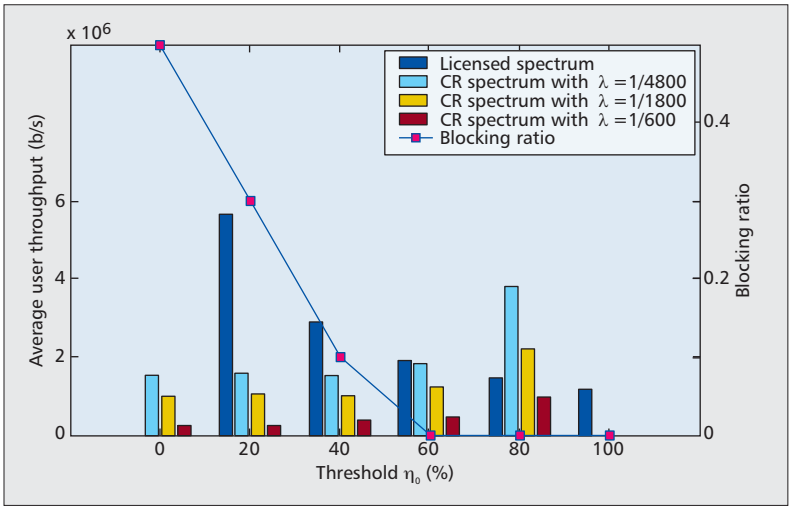


Figure 5. Average unblocked user throughput on licensed spectrum and cognitive spectrum.

rate λ and μ, respectively. Moreover, all users have CR capability, and those in idle state are activated and complete the communication following two independent Poisson processes with a mean rate of λ′ and μ′, respectively.

The blocking ratio of the conventional LTE and LTE-CR networks is shown in Fig. 4 as a function of the arrival rate of PUs. When all channels are fully loaded and there are no idle channels, new activated users are blocked. Blocking ratio is the ratio of blocking user numbers to the total activated user numbers. Not surprisingly, for any given LTE users activated rate λ′, the blocking ratio of LTE-CR is reduced compared to that of LTE, thanks to the help of TV channels. However, the improvement becomes smaller as the PU arrival rate λ increases since less TV channels are available. It can be expected that when the number of TV channels increases, the performance of LTE-CR can be further improved. Moreover, for a given PU arrival rate λ, the blocking ratio of both conventional LTE and LTE-CR increases with the LTE user activated rate λ′, with LTE-CR still outperforming LTE. Hence, the larger the user activated rate λ′, the busier the LTE network, and the higher the blocking ratio for both LTE and LTE-CR.

Figures 5 and 6 illustrate the impact of the load threshold η_0 on the average user throughput in LTE-CR using the SIB-CPC-based backward-compatible and load-aware network access. Assume there are 20 active users and one TV channel. No handoff scheme is applied, and users interrupted by PUs stay and wait for PU departure to continue data transmission. The average unblocked user throughput distribution on licensed and cognitive spectrum is shown in Fig. 5 together with the blocking ratio. When the threshold η_0 is zero, LTE-CR is always activated, and all 20 users are trying to use the cognitive spectrum while leave the licensed spectrum blank. Since the 5 MHz cognitive spectrum can support at most 10 users, the blocking ratio could be as high as 50 percent. As the threshold η_0 increases, more and more users stay in the licensed spectrum, and fewer users are offloaded to the cognitive one. Since the licensed spectrum with 10 MHz could support 20 users, the blocking ratio is reduced to zero at high η_0. Moreover, the average user throughput of the unblocked users on the licensed spectrum decreases since the spectrum is shared by more users. On the contrary, when η_0 increases, the CR spectrum is shared with fewer users, so the average user throughput increases. The average user throughput, including both the unblocked and blocked users, is shown in Fig. 6. For any threshold value but 100 percent, the smaller the PU arrival rate λ, the more available the CR spectrum and the higher the average user throughput. When η_0 is 100 percent, the system is equivalent to the conventional LTE and the average user throughput is irrelevant to η_0. Moreover, for a given PU arrival rate λ, the average user throughput in LTE-CR with the threshold from 20 to 80 percent are almost the same since the available spectrum resource and the total user number is fixed. Combining Figs. 5 and 6, an appropriate selection of the threshold would be between 0.6 to 0.8, resulting in good average user throughput with near-zero blocking ratio.

Conclusions

Network access is important for LTE-CR cellular networks using cognitive spectra in standalone mode. Based on the investigation on network access schemes applied in existing cognitive cellular systems such as IEEE 802.22 and IEEE 1900.4, CPC-based network access has been shown to be more suitable for LTE-CR, which could provide fast network access and put no stringent requirements on terminals. Moreover, a CPC-based backward-compatible network access scheme is desirable to facilitate the application of CR in LTE. Hence, a SIB-CPC-based network access scheme that carries CPC information on current PDSCH has been designed for LTE-CR, featuring little standard effort. Moreover, load awareness has also been introduced so that LTE-CR is activated only when the system load is high. The performance of this SIB-CPC-based backward-compatible and load-aware network access has been evaluated via simulations. It can

be seen that the blocking ratio of LTE-CR can be reduced notably compared with that of LTE. Moreover, by selecting an appropriate load threshold to activate LTE-CR, the average user throughput of LTE can be improved with near-zero blocking ratio by offloading users to complementary cognitive spectra.

ACKNOWLEDGMENTS

This work was supported by the national 863 program of China (No. 2014AA01A703), the National Natural Science Foundation of China (No. 61201231), Beijing Natural Science Foundation Project (No. 61222103) and the New Technology Star Plan of Beijing (No. xx2013052).

REFERENCES

[1] Z. Zhang, K. Long, and J. Wang, "Self-Organization Paradigms and Optimization Approaches for Cognitive Radio Technologies: a Survey," *IEEE Wireless Commun.*, Apr. 2013, pp. 36-42.

[2] T. Luan, F. Gao, and X. Zhang, "Joint Resource Scheduling for Relay-Assisted Broadband Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, Sept. 2012, pp. 3090–100.

[3] C. Cordeiro *et al.*, "IEEE 802.22: the First Worldwide Wireless Standard Based on Cognitive Radios," *IEEE DySPAN '05*, Nov. 2005, pp. 328–37.

[4] S. Buljore *et al.*, "Architecture and Enablers for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks: The IEEE 1900.4 Working Group," *IEEE Commun. Mag.*, vol. 47, no. 1, Jan. 2009, pp. 122–29.

[5] G. Dimitrakopoulos, P. Demestichas, and W. Koenig, "Introduction of Cognitive Systems in the Wireless World — Research Achievements and Future Challenges for End-to-End Efficiency," *Future Network and Mobile Summit '10*, June 2010, pp. 1–9.

[6] RP-140481, "New SID: Study on Licensed-Assisted Access Using LTE," Mar. 2014.

[7] RP-140060, "Summary of a Workshop on LTE in Unlicensed Spectrum," Mar. 2014.

[8] RP-140057, "On the Primacy of Licensed Spectrum in Relation to the Proposal of Using LTE for a Licensed-Assisted Access to Unlicensed Spectrum," Mar. 2014.

[9] 3GPP TS 36.304, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Procedures in Idle Mode," v. 11.0.0, June 2012.

[10] Z. Yan, G. Sun, and X. Wang, "A Novel Initial Cell Search Scheme in TD-LTE," *IEEE VTC-Spring '11*, May 2011, pp. 1–5.

[11] ITU-R Doc. 5A/788-E, "Annex 9-Working Document towards a Preliminary Draft New Report ITU-R [LMS.CRS2]-Cognitive Radio Systems [(CRS) Applications] in the Land Mobile Service," Working Party 5A, Nov. 2011.

[12] Q. Zhang *et al.*, "A Novel Mesh Division Scheme Using Cognitive Pilot Channel in Cognitive Radio Environment," *IEEE VTC-Fall '09*, Sept. 2009, pp. 1–6.

[13] J. Perez-Romero *et al.*, "A Novel On-Demand Cognitive Pilot Channel Enabling Dynamic Spectrum Allocation," *IEEE DySPAN '07*, Apr. 2007, pp. 46–54.

[14] H. Sun *et al.*, "Wideband Spectrum Sensing for Cognitive Radio Networks: A Survey," *IEEE Wireless Commun.*, Apr. 2013, pp. 74–81.

[15] J. Shen *et al.*, "3GPP Long Term Evolution (LTE): Principle and System Design," *Posts & Telecom Press*, Nov. 2008. (Chinese)

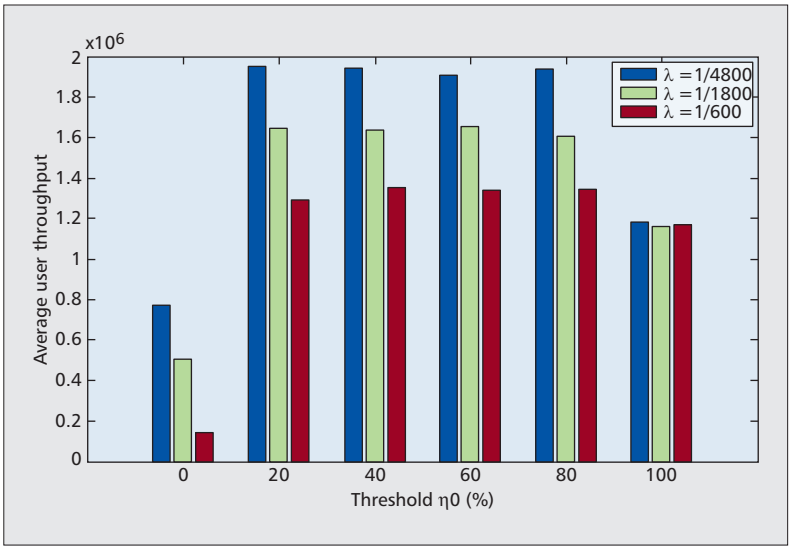


Figure 6. Average user (including blocked and unblocked) throughput in LTE-CR.

BIOGRAPHIES

LING LIU (liuling@ict.ac.cn) received her B.S. degree in communication engineering from Nanchang University in 2012. She is currently a Ph.D. candidate at the University of the Chinese Academy of Sciences. She serves as a reviewer for multiple international journals and conferences. Her research focuses on cognitive radio, dynamic spectrum sharing, and interference and resource management in heterogeneous networks.

YIQING ZHOU (zhouyiqing@ict.ac.cn) received her Ph.D. degree from the University of Hong Kong in 2004. Now she is a professor at the Wireless Center, ICT/CAS. She is or has been an Associate/Guest Editor for *IEEE TVT*, *IEEE JSAC*, *WCMC*, *ETT*, and *JCST*; TPC Co-Chair of ChinaCom '12, Symposium Co-Chair of IEEE ICC '15 and ICC '14, and Workshop Co-Chair of IEEE GLOBECOM '11. She received the Best Paper Award from IEEE WCNC '13 and ICCS '14, and a Top 15 Editor Award from IEEE TVT in 2014.

LIN TIAN (tianlindd@ict.ac.cn) is an associate professor at the Wireless Communication Research Center of ICT/CAS. She received her Ph.D. degree from ICT/CAS in April 2012. Her research interests include wireless resource management and multimedia multicast schemes in next-generation mobile communication systems. She was the Symposium Co-Chair of ChinaCom '13 and Publication Chair of ChinaCom '12. She has also served as a reviewer for a number of refereed journals and international conferences.

JINGLIN SHI (sjl@ict.ac.cn) is the director of the Wireless Communication Research Center of ICT/CAS. He is also a visiting professor at Beijing University of Posts and Telecommunications, the University of Sydney, the University of Wollongong, and Macquarie University. His research interests include wireless communications system architecture and management, wireless signal processing theory, and wireless communications baseband processor design. As a team leader, he successfully led the development of TD-SCDMA, WiMAX, and LTE protocol stack systems.

A Game-Theoretic Perspective on Self-Organizing Optimization for Cognitive Small Cells

Yuhua Xu, Jinlong Wang, Qihui Wu, Zhiyong Du, Liang Shen, and Alagan Anpalagan

ABSTRACT

In this article, we investigate self-organizing optimization for cognitive small cells (CSCs), which have the ability to sense the environment, learn from historical information, make intelligent decisions, and adjust their operational parameters. By exploring the inherent features, some fundamental challenges for self-organizing optimization in CSCs are presented and discussed. Specifically, the dense and random deployment of CSCs brings about some new challenges in terms of scalability and adaptation; furthermore, the uncertain, dynamic, and incomplete information constraints also impose some new challenges in terms of convergence and robustness. For providing better service to users and improving resource utilization, four requirements for self-organizing optimization in CSCs are presented and discussed. Following the attractive fact that the decisions in game-theoretic models are exactly coincident with those in self-organizing optimization (i.e., distributed and autonomous), we establish a framework of game-theoretic solutions for self-organizing optimization in CSCs and propose some featured game models. Specifically, their basic models are presented, some examples are discussed, and future research directions are given.

INTRODUCTION

Small cells have been regarded as a promising approach to meet the increasing demand of cellular network capacity. In comparison to macrocells, low-cost small cells operating with low power and short range offer a significant capacity gain due to spatial reuse of spectrum. Researchers in the community have realized that enabling cognitive ability into small cells, which is referred to as cognitive small cells (CSCs) [1], would further improve resource utilization. Similar to cognitive radio, CSCs are able to sense the environment, learn from historical information, make intelligent decisions, and adjust their operational parameters.

It is expected that small cells are to be *densely* deployed in the near future. Furthermore, small cells may be deployed by mobile operators, enterprises, or households, which means that they would operate in a self-organized, dynamic, and distributed manner. Thus, resource optimization problems for small cells (e.g., spectrum sharing, carrier selection and power control, interference management, and offloading mechanism) cannot be solved in a centralized manner since it results in heavy communication overhead and cannot adapt to a dynamic environment. As a result, it is important and timely to develop self-organizing optimization approaches for CSCs.

In this article, by exploring the inherent features of CSCs, we first discuss and analyze some fundamental challenges and requirements for self-organizing optimization in CSCs. Following the attractive advantages of game-theoretic models for self-organizing optimization, we propose some featured game-theoretic solutions. It should be pointed out that there are also some other useful approaches for self-organizing optimization in distributed wireless networks, for example, the swarm intelligence inspired evolutionary algorithms [2]. The reasons for using game-theoretic solutions are:

- The interactions among multiple decision makers can be well modeled and analyzed.
- The outcome of the game is predicable; hence, the system performance can be improved by manipulating the utility function and the action update rule of each decision maker.

Game-theoretic models have been investigated extensively in wireless communications, and there are some preliminary game-theoretic solutions for CSCs, such as reinforcement learning with logit equilibrium for power control [3], a hierarchical dynamic game approach for spectrum sharing and service selection [4], and an evolutionary game for self-organized resource allocation [5]. The presented models in this article mainly address the inherent features, fundamental requirements, and challenges of CSCs and hence differ significantly from previous ones seen in the literature. In fact, the main objective

Yuhua Xu, Jinlong Wang, Qihui Wu, Zhiyong Du, and Liang Shen are with PLA University of Science and Technology.

Alagan Anpalagan is with Ryerson University.

0163-6804/15/$25.00 © 2015 IEEE

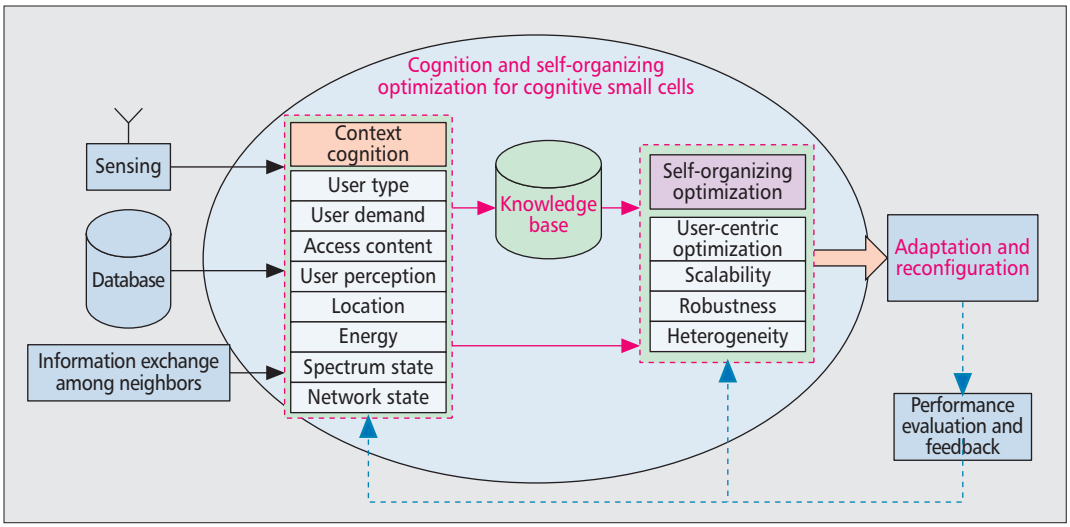


Figure 1. The paradigm of cognition functionality and self-organizing optimization in CSCs.

We argue that the term "cognitive" in CSCs is not limited to observing the environment and acquiring contextual information. Instead, it should include having high-level intelligence. To achieve such high-level intelligence for CSCs, the most promising way is to realize knowledge discovery from the contextual information.

of this article is to propose and discuss the featured game-theoretic models suitable for CSCs.

The rest of this article is organized as follows. The cognition functionality for CSCs is presented. Some fundamental challenges and requirements for CSCs are discussed. Some featured game-theoretic models for self-organizing optimization in CSCs are presented, and future research directions are given. Finally, we provide concluding remarks.

Cognition and Self-Organizing Optimization for Cognitive Small Cells

We first present the cognition functionality for CSCs, which is the basis of self-organizing optimization. The cognition functionality in cognitive radio is mainly concerned with acquiring spectrum availability information (i.e., sensing and identifying spectrum opportunities in time, frequency, and space domains). To capture the complex environment and network state, the cognition functionality in CSCs is extended to explore multidimensional information. Such multidimensional information is referred to as *contextual information*, which is used to identify an object of interest.

As illustrated in Fig. 1, the target contextual information includes user type, user demand, access content, user perception, location, energy, network state, and spectrum state. For presentation, we briefly illustrate the above contextual information. The user type represents the hardware category (tablet, phone, or laptop). Access content is related to a specific application, such as browsing breaking news or downloading an app. User perception is related to the service quality experienced by the user, while location is where the small cell is (e.g., indoors or outdoors). Energy is related to the power for wireless transmission and cooling. Network state is related to the current network deployment of small cells, and spectrum state is related to the spectrum availability. Technically, the above contextual information has great impact on the resource allocation schemes, which are discussed later.

As shown in Fig. 1, the contextual information in CSCs can be obtained by the following three methods: sensing, database, and information exchange among neighbors.

Sensing: With the help of cognitive radio technologies, CSCs perform sensing to obtain useful information. For example, the spectrum occupancy state can be obtained by energy detection or feature detection (e.g., pilot, modulation type, cyclic prefixes, and cyclostationarity). Sensing is real-time but consumes resources including time, energy, and bandwidth.

Database: The database approach is a powerful tool to provide useful information for CSCs. For example, a spectrum database has recently been developed to provide spectrum occupancy state for a particular region, through which the CSCs can request location-dependent spectrum availability information. Compared to sensing, the database approach is more efficient but is not real-time.

Information exchange among neighbors: Since the CSCs are connected to the core network via cable or optical fiber, information exchange among CSCs is feasible. However, note that local information exchange between neighbors is more desirable since global information exchange leads to heavy communication and signaling overhead.

We argue that the term *cognitive* in CSCs is not limited to observing the environment and acquiring contextual information. Instead, it should include having high-level intelligence. To achieve such high-level intelligence for CSCs, the most promising way is to realize knowledge discovery from the contextual information. Generally speaking, knowledge is a broad concept including general principles and natural laws. Taking the spectrum dynamics as an example, the probability θ that a particular band is occupied by macrocells during 01:00 a.m. to 04:00 a.m. is very small (e.g., $\theta = 0.05$) is viewed as knowledge in CSCs.

Based on the contextual information, the CSCs can build a knowledge base that contains useful knowledge for self-organizing optimization.

With the increase in the number of small cells, how would the self-organizing optimization solutions scale up? This is the first basic issue of dense deployment in CSCs. In addition, since the decisions of CSCs are interactive, addressing the complicated interactions among densely deployed CSCs is another important issue.

Challenges and Requirements for Self-Organizing Optimization in Cognitive Small Cells

In this section, by exploring the inherent features of CSCs, we briefly discuss some fundamental challenges and requirements for self-organizing optimization in CSCs.

Technical Challenges

The technical challenges for self-organizing optimization in CSCs are discussed from the perspectives of network deployment and information constraints respectively.

First, from the perspective of network deployment of CSCs, two challenges arise in the following two aspects:

Scalability: With the increase in the number of small cells, how would the self-organizing optimization solutions scale up? This is the first basic issue of dense deployment in CSCs. In addition, since the decisions of CSCs are interactive, addressing the complicated interactions among densely deployed CSCs is another important issue.

Random deployment: Small cells are deployed by different entities (e.g., mobile operators, enterprises, or households). In addition, they may be inactive if there is no serving client. As a result, the deployment of small cells is always random and dynamic. Thus, it is important for self-organizing optimization solutions to behave in random and dynamic environments.

Second, it is known that information is key to optimization problems, and the challenges related to information arising in the CSCs are listed below.

Uncertainty: The observed information may not be the same as the true information. A well-known example is that the sensed spectrum states are always imperfect due to the corruption of noise.

Dynamic: The observed information is time-varying, and the dynamic changes are not determinate. For example, the network and spectrum states may change from time to time, and the set of active CSCs may also change from time to time. Furthermore, the network and spectrum states in each decision period are random, the set of active CSCs are random, and their demands are random as well.

Incomplete: Due to the constraints in hardware and resource consumption, each CSC only has partial information about the environment; furthermore, it only has information on its neighboring CSCs (in some extreme scenarios, it has no information on others). In addition, a CSC does not know the total number of small cells in any systems, not to mention the active ones.

Due to the above technical challenges, it is seen that the task of resource optimization in CSCs is hard to solve even in a centralized manner, not to mention in a distributed and self-organizing manner.

Requirements for Self-Organizing Optimization

We list some fundamental requirements for self-organizing optimization in CSCs. Specifically, these requirements are for user service, network deployment (architecture), and optimization methodology. As shown in Fig. 1, based on the contextual information and knowledge, some self-organizing optimization approaches can be applied to resource allocation in CSCs. By employing their inherent features, we discuss some featured requirements of self-organizing optimization in CSCs, which mainly include user-centric optimization, scalability, robustness, and heterogeneity.

First, it should shift from throughput-oriented optimization to user-centric optimization. Traditionally, resource optimization schemes in wireless systems are throughput-oriented, with the objective to maximize throughput/capacity or minimize delay. However, it is now realized that throughput-oriented schemes cannot provide satisfactory service for users. In future mobile communication systems, there is an increasing trend to develop user-centric optimization schemes rather than throughput-oriented schemes. The underlying reasons are twofold:

- Eventually, the purpose of (wireless) communication is to serve end users. Thus, the contextual information of users (e.g., their locations, demands, access contents, and energy) should be taken into account not only at high layers but also at the physical (PHY) and medium access control (MAC) layers for optimization
- It is realized that mobile (cellular) systems have migrated toward data and Internet services. In particular, multimedia service delivery through cellular systems (e.g., watching online video) is becoming common. For this kind of service, people may not care about the specific volume of allocated resources, but sensitively react to the perceived service quality, which is known as quality of experience (QoE) [6]. This means that user perception should also be taken into account in self-organizing optimization.

Second, it should admit scalability and address network density. As stated before, it is expected that CSCs will be densely deployed in large numbers. A consequence is that the resource optimization for dense deployment is completely different than that in a sparse environment. Thus, the self-organized optimization schemes should scale up in dense CSCs. In addition, density creates congestion and interference among CSCs, which implies that efficient congestion control and interference mitigation approaches should be developed.

Third, it should be robust to the dynamic environment. As discussed before, there are several random and dynamic factors in CSCs (e.g., the spectrum availability is dynamic; the CSCs switch between active and inactive randomly). Moreover, the observed information may be corrupted by noise. Thus, self-organizing optimization solutions should be robust to address the randomness, dynamics, and uncertainty in CSCs.

Last, it should address the hierarchical decision making in CSCs. There are always heterogeneous cells with overlapping coverage in future wireless

systems, that is, macrocells and small cells. In such hierarchical networks, the cells at different layers have different priority and utility functions. Hence, it involves heterogeneous decision makers. However, traditional self-organizing optimization solutions are mainly for homogeneous decision makers. Thus, it is important to develop new hierarchical self-organizing solutions for CSCs.

Game-Theoretic Self-Organizing Optimization for Cognitive Small Cells

Game theory [7] is an applied mathematic tool to model and analyze mutual interactions in multiuser decision systems. Generally, a game model consists of a set of players, a set of available actions of each player, and a utility function that maps the action profiles of all the players into a real value. There are two major branches of game-theoretic models: non-cooperative games and cooperative games. From a high-level comparison perspective, players in a non-cooperative game make rational decisions to maximize their individual utility functions, while players in a cooperative game are grouped together according to an enforceable agreement for payoff allocation. In a non-cooperative game, the commonly used solution concepts are Nash equilibrium (NE) and correlated equilibrium.

Researchers began to apply game-theoretic models to wireless communications a decade ago; nowadays, it is regarded as a powerful tool for wireless resource allocation optimization, such as power control, spectrum access, network selection, spectrum auction and trading, and incentive mechanism design. The decisions of the players in (non-cooperative) game-theoretic models are distributed and autonomous, which is an exact coincidence with those in self-organizing optimization. Thus, a game-theoretic approach is important to achieve self-organizing optimization in CSCs [8].

Framework of Game-Theoretic Self-Organizing Optimization

To cope with the technical challenges in CSCs — dense and dynamic deployment, and uncertain, dynamic, and incomplete information constraints — we propose a framework of game-theoretic self-optimizing optimization, which is shown in Fig. 2. It is noted that there are two key steps:

- Game formulation and analysis
- Design of multiuser learning algorithm

On one hand, the stable solutions are the inherent properties of game-theoretic models, and not relevant to the learning algorithms. On the other hand, except for the utility function in game-theoretic models, the uncertain, dynamic, and incomplete information constraints have great impact on the convergence and performance of learning algorithms.

Game Formulation and Analysis: For game formulation, one needs to first identify the player and available action set, and define suitable utility functions for the players. For CSCs, the player may be a single entity (e.g., small base station or user equipment) or a collection of multiple entities (e.g., a cluster consisting of multiple nearby small cells). The available action set can be regarded as a combination of multiple optimization variables. Defining utility function is key to game formulation since it eventually determines the properties and performance of the game-theoretic models.

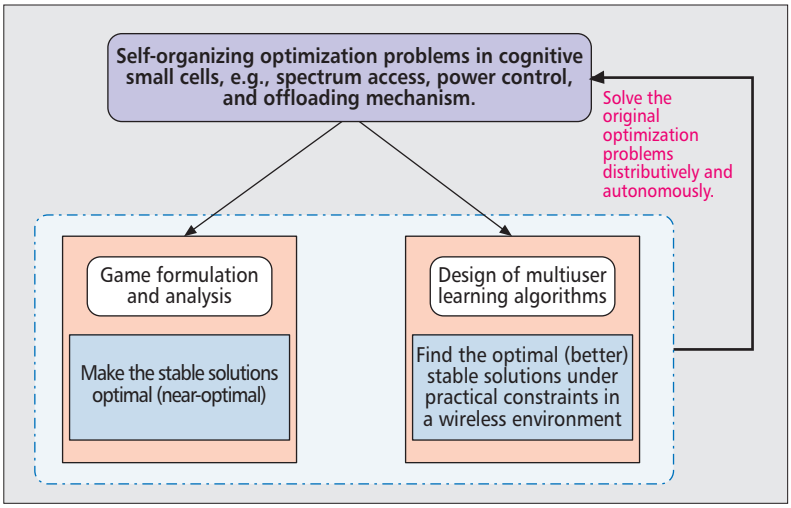


Figure 2. The proposed framework of game-theoretic solutions for self-organizing optimization in CSCs.

The most efficient game-theoretic model used in wireless networks is potential game [9], in which there is a potential function such that the change in the utility function caused by the unilateral action change of an arbitrary player has the same trend with that in the potential function, both increasing and decreasing. Potential game has at least one pure strategy NE, and all NE points are global or local maxima of the potential function. Thus, the NE solutions are desirable if the potential function is related to the original optimization objective. Furthermore, to ensure that the stable solutions of game-theoretic models are optimal (near-optimal), another efficient method is to define the utility function as the received payoff minus the cost of using the amount of a particular resource.

Design of Multiuser Learning Algorithms: Identifying the stable solutions of game-theoretic models is one thing, but finding them is a different thing. This issue, however, was underestimated in previous studies. In pure game theory, players can monitor the environment and other players, which means that they have perfect information about the actions and payoffs of other players. As discussed above, this assumption does not hold in CSCs. With the cognition functionality of CSCs, players need to observe the results of multiuser interactions (e.g., interference, collision, and competition), learn useful information from limited feedback, and then adjust their behavior toward some desirable solutions. In the context of game optimization, the objective of multiuser learning is to converge to a stable solution with good performance.

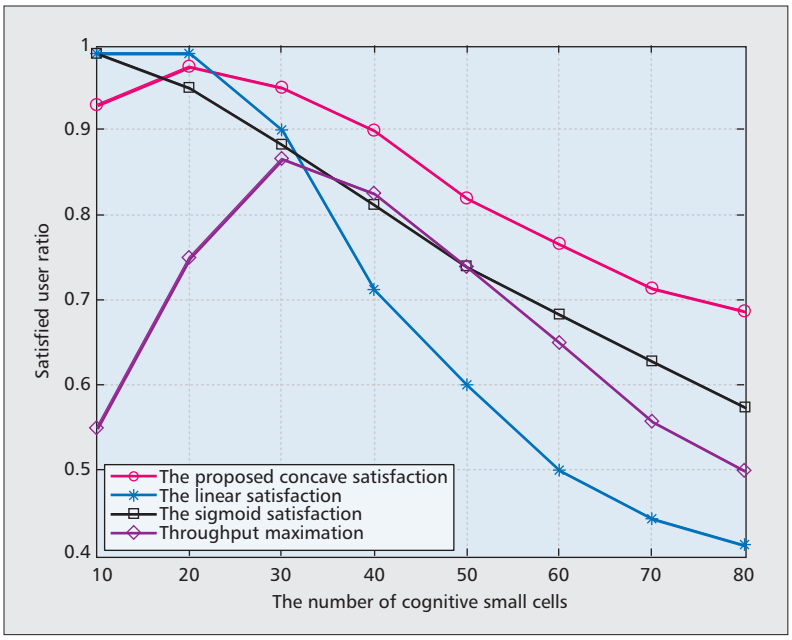


Figure 3. The comparison results of satisfied user ratio for different satisfaction utility functions.

Denote $a_n(k)$ as the action of player n in the kth iteration, and $a_{-n}(k)$ as the action profile of all other players except n. Due to the interactions (interference, congestion, or competition) among players, the received payoff $r_n(k)$ of each player is jointly determined by the action profile of all players, and it may be deterministic or random. Generally, the players update their actions based on the current action-payoff information $\{a_n(k), a_{-n}(k); r_n(k), r_{-n}(k)$. Thus, the system evolution can be described as $\{a_n(k), a_{-n}(k)\} \rightarrow \{r_n(k), r_{-n}(k)\} \rightarrow \{a_n(k+1), a_{-n}(k+1)\}$, and the objective is to converge to a stable action profile that maximizes system utility.

The uncertain, dynamic, and incomplete information constraints in CSCs may pose some new challenges. Specifically:

- A player does not know the information about all other players (i.e., $a_{-n}(k)$ and $r_{-n}(k)$ are unknown).
- The received payoff $r_n(k)$ may be random and time-varying.

Thus, the update rule needs to be carefully designed to guarantee the convergence toward desirable solutions. When local information among neighboring players is available, it is desirable to develop partially uncoupled learning algorithms based on the partial action-payoff information $\{a_n(k), a_{J_n}(k); r_n(k), r_{J_n}(k)\}$, where $a_{J_n}(k)$ and $r_{J_n}(k)$ are the action-payoff information of the neighboring players. In some extreme scenarios with no information exchange available, one needs to develop fully uncoupled learning algorithms based on the individual action-payoff information $\{a_n(k), r_n(k)\}$. There are some useful partially coupled learning algorithms, such as local altruistic behavior with spatial adaptive play [10], and fully uncoupled learning algorithms (e.g., stochastic learning automata [11]) that can be applied to self-organizing optimization in CSCs.

Some Featured Game-Theoretic Models for CSCs

In methodology, previous game-theoretic models in wireless communications can be applied to CSCs. However, most previous game-theoretic solutions mainly focused on analyzing the properties of game-theoretic models in ideal scenarios, and did not take into account the challenges and requirements in CSCs. In this subsection, we propose some featured game modes for CSCs. Due to the low-power, the transmission of a small cell only affects its neighbors; as a result, graphical games [10] are appropriate for small cell networks.

Demand-Aware Game: Most existing resource optimization approaches have mainly focused on maximizing resource utilization, while ignoring the actual demand of users. In future CSCs, demand-aware design and decision are more desirable. To include user demand in the resource optimization problems, a useful method is to map the allocated resource to the user satisfaction utility. Specifically, denote user n's demand as d_n and the allocated resource as r_n; then its satisfaction utility can be expressed as $s_n(r_n, d_n)$.

Generally, there are two kinds of satisfaction functions in the literature:

- Linear satisfaction: The satisfactory utility is determined by r_n/d_n if $r_n \le d_n$, and is equal to one otherwise.
- Sigmoid satisfaction: The satisfactory utility function is generally determined by

$$s_n(r_n, d_n) = \frac{1}{1+e^{-c(r_n-d_n)}},$$

where c is used to adjust the slope of the satisfaction utility curve around the user demand d_n with different types of traffic.

In particular, real-time traffic such as online video is sensitive to acquired resources and has strict performance requirements, which corresponds to a large value of c, while non-real-time traffic such as email or file transfer is less sensitive, which corresponds to a small value of c. The linear and sigmoid satisfaction functions have been well investigated in previous game-theoretic wireless resource optimization problems. As the satisfaction utility function is strictly increasing, each user proceeds to compete for wireless resource even if the obtained resource is larger than the demand, which would decrease the satisfaction of others. However, this drawback has not been addressed in previous work.

To improve network satisfaction, an efficient approach is to prevent users competing for extra resources when they are satisfied and to decrease the satisfaction utility when it occupies additional resources. Based on this intuition, the concave satisfaction utilities may be more suitable for multiuser communication networks. An example of concave satisfaction utilities is given by

$$\left(\frac{2\sqrt{r_n d_n}}{d_n + r_n}\right)^{\alpha},$$

where α is used to adjust the slope of the utility curve for different types of traffic. For illustration, we consider the problem of distributed

spectrum access for CSCs. Specifically, the CSCs are randomly located in a region of 100 m × 100 m, and the sensing-based spectrum access protocol proposed in [1] was applied. The problem of distributed spectrum access is formulated as a graphical game, and the stochastic learning automata [11] is applied. Different typical applications, such as G.711PCM, WMV, AVI/RM, Flash, and H.264, are considered in the simulation. The comparison results are presented in Fig. 3. It is noted that with the proposed satisfaction function, the satisfaction user ratio is largely improved. In particular, as the network scales up, the throughput gain becomes significant.

Discrete-QoE-Aware Game: Eventually, the purpose of wireless communications is to serve people. Thus, the perception by people, QoE, should be included in the game formulation. Unlike the satisfaction function, which is characterized by continuous and real values, the perception of people is generally subjective and discrete. For example, a person may feel "Excellent," "Good," "Fair," "Poor," and "Bad" by the the mean opinion score method [6]. Compared to a traditional continuous optimization game, an interesting result of the discrete-QoE-aware game is the expansion of NE, which is shown in Fig. 4.

In a traditional continuous optimization game, users maximize their throughput as there is an inherent principle: larger throughput is always better. On the contrary, a user in discrete-QoE-aware games does not always maximize its throughput unless its QoE level can be improved, say, from "Poor" to "Fair." Thus, it can be expected that a discrete-QoE-aware game would improve the network QoE.

To further show the benefit of a discrete-QoE-aware game, we consider the problem of distributed user association in Long Term Evolution-Advanced (LTE-A) small cell networks [12]. For users located in the overlapping areas, there are multiple small cell access points (SAPs) available, and the users need to choose one with which to associate. Consider three types of video call users using Skype [12]:

- The first one is the group video call with the required minimal throughput R_m = 512 kb/s and the recommended throughput R_c = 2 Mb/s.
- The second one is the high definition video calling with R_m = 1.2 Mb/s and R_c = 1.5 Mb/s.
- The third one is the general video calling user with R_m = 128 kb/s and R_c = 500 kb/s.

Each user falls into one of the above three types with equal probability. It is believed that the minimal throughput only supports the basic user demand ("Poor"), while the recommended throughput offers sufficiently good user experience ("Good"). With the method proposed in [12], the throughput thresholds for other QoE levels ("Excellent," "Fair," and "Bad") can be obtained accordingly.

Considering a network with 78 users that can access only one SAP, and 20 users located in the overlapping regions of neighboring CSCs, the comparison results of the number of users at different QoE levels are shown in Fig. 5. It is seen that the discrete-QoE-aware game outperforms the continuous optimization game. Specifically, with the discrete-QoE-aware game, 12 users are in "Excellent," one in "Good," two in "Fair," and five in "Bad"; with the continuous optimization game, 11 users are in "Excellent," none in "Good," three in "Fair," and five in "Bad". Also, it is noted that the throughput maximization approach (the user demand is neglected) achieves poor network QoE. This result validates the superiority of the discrete-QoE-aware game.

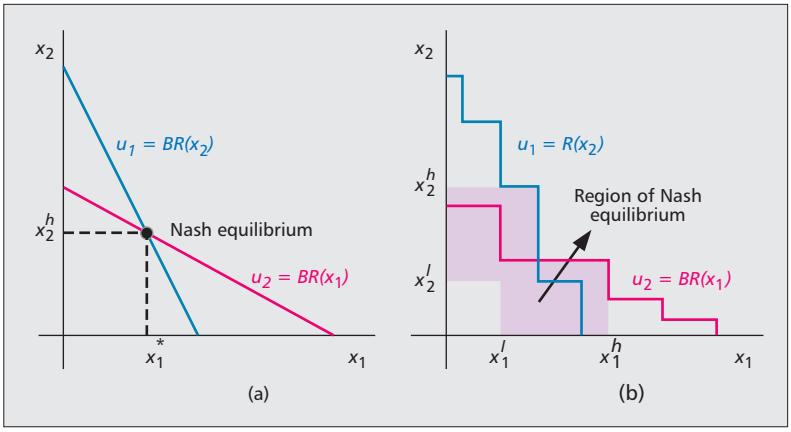


Figure 4. An illustrative expansion of NE of a discrete-QoE-aware game. $BR(x_2)$ ($BR(x_1)$) denotes the best response curve of player 1 (player 2) with respect to the decision variable of the other player: a) an illustrative diagram of NE for games with continuous utility function; the intersection point (x_1^*, x_2^*) is NE; b) an illustrative diagram of NE for QoE-aware game with non-continuous utility function; due to the discontinuous feature of the QoE-aware game, NE is expanded to the shadow region.

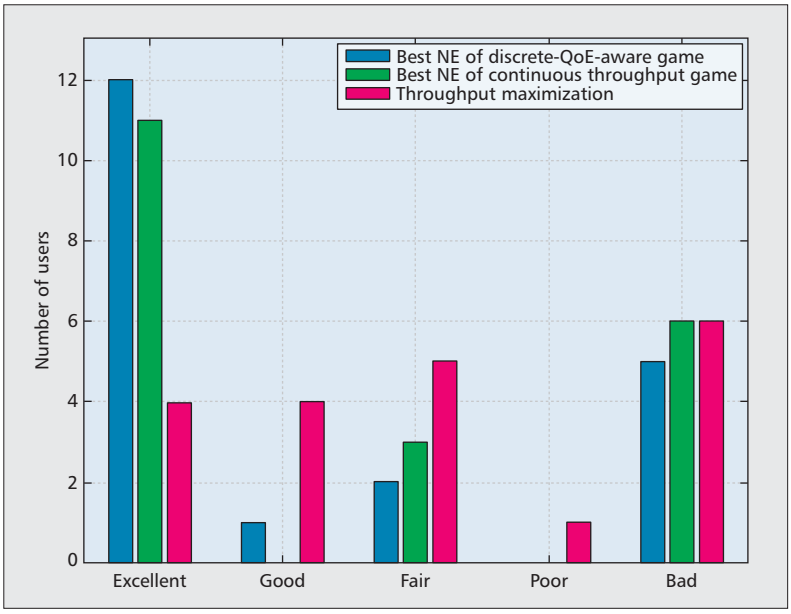


Figure 5. The number of users in different QoE levels of different solutions.

Hierarchical Game: As stated before, CSCs would interact with macrocells for dynamic spectrum sharing and mobile offloading. While originally studied in the economic context of duopolies in which one company has the power to act before the other companies, the Stackelberg game, which is an important kind of hierar-

chical game, is suitable for systems that contain a natural hierarchy. Therefore, to address the hierarchical decision making between macrocells (always acting as a leader) and CSCs (always acting as followers), the Stackelberg game is becoming a useful tool [4].

In addition, following the idea of "divide-and-conquer," a hierarchical game can also be used to address the dense deployment of CSCs. In particular, in order to ease the challenges caused by the large number of participants, we can create hierarchy to transform the large-scale optimization problem into several layers of sequential sub-problems. To achieve this, a useful method is clustering. An example of creating hierarchy to use cluster-based hierarchical game in large-scale CSCs is shown in Fig. 6. In Fig. 6a, the network topology and the interference relationship are presented. In Fig. 6b, the neighboring small cells distributively form two disjoint clusters, with cell 3 and cell 6 serving as the cluster headers, respectively. In Fig. 6c, in the upper layer, the headers compete for resources with each other, aiming to maximize the aggregate utility of the cluster; in the lower layer, the cluster members compete with other members, under the policies imposed by the header. In Fig. 6d, since different clusters behave independently, there may be interference between neighboring clusters (e.g., cells 4 and 5 still interfere with each other). Thus, the interfering cells further mitigate mutual interference via distributed learning, (e.g., Q-learning). With the proposed cluster-based hierarchical structure, the self-organizing optimization in large-scale networks can then be solved with moderate computational complexity.

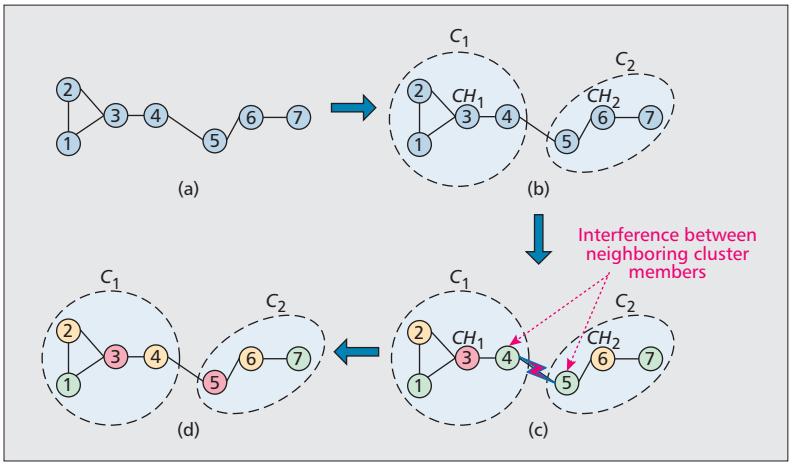


Figure 6. An example of using "divide-and-conquer" to use cluster-based hierarchical game in large-scale CSCs. In c) and d), the colors represent the channels chosen by the small cells.

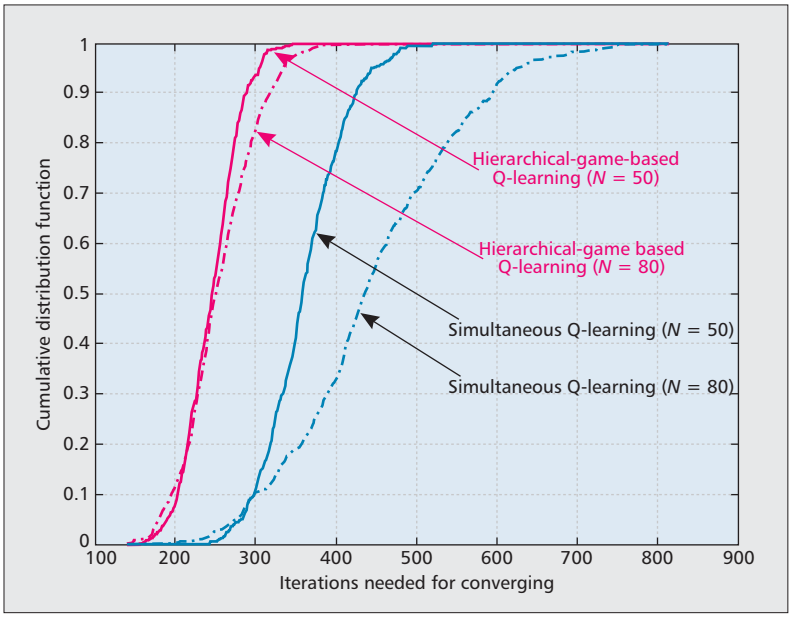


Figure 7. The convergence speed comparison between hierarchical-game-based Q-learning and simultaneous Q-learning.

We compare the computational complexity between the proposed cluster-based hierarchical game and the simultaneous Q-learning approach [13], in which all cells performs Q-learning simultaneously. The achievable network throughput of two approaches is almost the same. The cumulative distribution function (CDF) of the iterations needed to converge is shown in Fig. 7. It is noted from the figure that for the same size network ($N = 50$ or $N = 80$), the iterations needed to converge with the hierarchical game Q-learning approach are significantly decreased. Furthermore, when the network scales up from $N = 50$ to $N = 80$, the convergence speed of the hierarchical-game-based Q-learning approach slightly decreases, while that of the simultaneous Q-learning approach is much decreased. This implies that the proposed hierarchical-game-based approach is especially suitable for dense and large-scale networks.

Robust Game: To capture the random and dynamic behavior in CSCs, a robust game is a good candidate. Specifically, the utility function in robust games is defined over statistics [14], such as expectation or other high-order statistics. In the following, we present a robust spectrum access game for CSCs as an illustrative example.

Consider a distributed CSC network operating in the TV white space. Each cognitive SAP inquires about the spectrum availability from the geo-location database, which specifies the available channel set and the maximum allowable transmission power for each cell. To capture the dynamic cell load in practical applications, we consider a network with a varying number of active cells. Specifically, it is assumed that each cell performs spectrum access with probability λ_n, $0 < \lambda_n \leq 1$, in each decision period. Note that such a model captures general kinds of dynamics in wireless networks; for example, a cell becomes active only when it has data to transmit and inactive when there is no transmission demand. Also, it can be regarded as an abstraction of the dynamic cell loading, that is, the cell active probability corresponds to the probability of a non-empty loading buffer. Note that the active cell set in each period is not deterministic and randomly changes from period to period. Also, a cell does not know the active probabilities of other cells.

To address the dynamic and random deployment of CSCs, a robust spectrum access game can be formulated in which the utility function of a CSC is defined as the expected Shannon capacity over all possible active cell sets. The

game can be proved to be a potential game [9]; hence, the distributed learning automata algorithm [11] can be applied to converge to NE points in the dynamic environment. Taking a network with nine CSCs as an illustrative example, the throughput performance comparison results are shown in Fig. 8. The optimum is obtained using the exhaustive search method in a centralized manner, by assuming that there is a genie that knows all required information. The best and worst NE is obtained using the best response algorithm in a distributed manner, by assuming that information exchange among neighboring cells is available. Some important results can be observed:

- The best NE is almost the same as the optimal one, while the throughput gap between the worst NE and the optimum is also trivial, which validates the effectiveness of the formulated robust spectrum game.
- The achievable throughput of the distributed learning automata is very close to the optimal one.

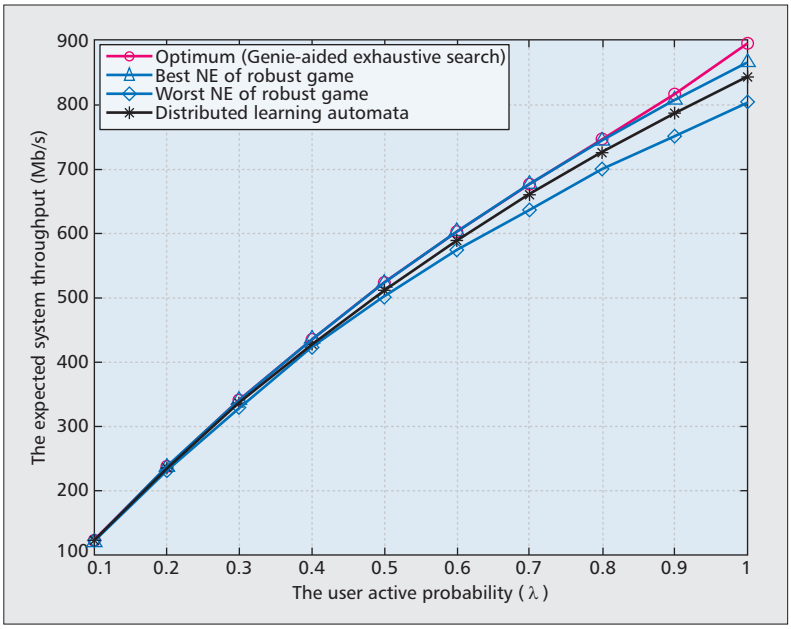


Figure 8. The expected Shannon capacity when varying the active probabilities of the cells.

Content-Aware Game: As legacy cellular systems have migrated toward data and Internet services, taking into account the access content in the self-organizing optimization would enjoy content gain. For Internet traffic, it has been shown in [15] that a relatively small portion of the access items accounts for a vast fraction of the information accesses, and Zipf's law can be used to determine the occurrence frequency of the access items, given the content rank, the content pool size, and the characteristic curve of the access pattern. Nowadays, content caching has become a core technology for wireless cellular systems. Thus, it is reasonable to replicate significant portions of popular contents on the wireless caches. As a result, the search and access time for popular content is fast compared to that of unpopular content. The reason is that popular content can be accessed in wireless caches, while unpopular content is accessed from faraway servers. Therefore, the differences in access time of different contents will have a great impact on wireless resource allocation, and it is promising to explore content-aware game-theoretic solutions for CSCs, which would achieve better performance.

Comparative Summarization and Analysis

In comparison, the game-theoretic models for CSCs presented in this article differ from previous ones significantly. Specifically, they shift from throughput-oriented optimization to user-centric optimization (e.g., demand-aware game, discrete-QoE-aware game, and content-aware game), address the dense deployment of small cells (e.g., graphical game and hierarchical game), and cope with randomness and dynamics well (e.g., robust game). Although the research on game-theoretic self-organizing optimization for CSCs is in infancy, we believe that the presented game-theoretic models will draw great attention in the near future.

For a specific resource optimization problem in CSCs, one can choose a suitable game-theoretic model and a learning algorithm to construct a self-organizing optimization solution. However, it should be pointed out that a game-theoretic solution for CSCs is application-dependent, which means that the game-theoretic model and distributed learning algorithm should be carefully formulated and designed.

Future Research Direction

It is seen that game-theoretic solutions for self-organizing optimization in CSCs definitely have a beautiful and exciting future, although current research is still far away from the expected vision. We list some future research problems for game-theoretic models and learning procedures below.

- Develop or investigate new game-theoretic models for self-organizing optimization from social/biological behaviors. The rationale behind this is that in old days humans first self-organized and then evolved successfully with population growth. For example, motivated by local altruistic behavior in biological systems, a local altruistic game with each player maximizing its utility and the aggregate utilities of its neighbors was proposed to achieve global optimization via local information exchange [10]. The key to the design with this issue is to properly abstract and model the social/biological behaviors, which is interesting and challenging.
- It is noted that each kind of game presented mainly addresses a single aspect of challenges in CSCs. However, as can be expected, one may combine more than one game-theoretic model (e.g., robust discrete-QoE-aware game or Stackelberg graphical game) to address multiple aspects of challenges of CSCs simultaneously. Such combinations bring about new challenges since the game structure is completely changed.
- Design and analyze heterogeneous learning algorithms. In most existing studies, it is assumed that all the decision makers employ the same learning algorithm. However, this assumption is for academic research but not true in practical systems. In practice, the small cells may belong to

Knowledge can be viewed as high-level intelligence obtained from the contextual information, which is truly beneficial for decision-making. Thus, we should develop some new knowledge-assisted learning technologies to increase the converging speed and achieve better performance.

different holders, which may adopt different learning algorithms; in addition, even the small cells belonging to the same holder may have different processing ability and preference, and hence choose heterogeneous learning algorithms. Introducing heterogeneity into the learning procedure will change the convergence and asymptotic behavior, which needs to be further studied.

• Design knowledge-assisted learning algorithms. The common procedure in existing learning algorithms is to update the strategies based on the historical action-payoff information. It may take a long time to converge to stable solutions since the players need to explore all the possible actions. As shown in Fig. 1, knowledge can be viewed as high-level intelligence obtained from contextual information, which is truly beneficial for decision making. Thus, we should develop some new knowledge-assisted learning technologies to increase the convergence speed and achieve better performance.

Conclusion

In this article, we investigate self-organizing optimization for CSCs, which will play an important role in future cognitive cellular systems. By exploring the inherent features, some fundamental challenges and requirements for self-organizing optimization in CSCs are presented and discussed. Following the attractive advantages of game-theoretic models (i.e., distributed and autonomous decision making), a framework of game-theoretic solutions for self-organizing optimization in CSCs is established, and some featured game-theoretic models are proposed. Specifically, the basic game-theoretic models are presented, some insights are discussed, some examples are discussed, and future research directions are given.

Acknowledgment

This work was supported by the National Science Foundation of China under Grant No. 61401508 and No. 61172062.

References

[1] H. Elsawy, E. Hossain and D. I. Kim, "HetNets with Cognitive Small Cells: User Offloading and Distributed Channel Access Techniques," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 28–36.
[2] Z. Zhang *et al.*, "On Swarm Intelligence Inspired Self-Organized Networking: Its Bionin Mechanisms, Designing Principles and Optimization Approaches," *IEEE Commun. Surveys and Tutorials*, vol. 16, no. 1, Feb. 2014, pp. 513–37.
[3] M. Bennis *et al.*, "Self-Organization in Small Cell Networks: A Reinforcement Learning Approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, July 2013, pp. 3202–12.
[4] K. Zhu, E. Hossain, and D. Niyato, "Pricing, Spectrum Sharing, and Service Selection in Two-Tier Small Cell Networks: A Hierarchical Dynamic Game Approach," *IEEE Trans. Mobile Comp.*, vol. 13, no. 8, Aug. 2014, pp. 1843–56.
[5] P. Semasinghe, E. Hossain, and K. Zhu, "An Evolutionary Game for Distributed Resource Allocation in Self-Organizing Small Cells," *IEEE Trans. Mobile Comp.*, vol.14, no. 2, Jan. 2015, pp. 274–87.
[6] J. A. Hassan *et al.*, "Managing Quality of Experience for Wireless VOIP Using Noncooperative Games," *IEEE JSAC*, vol. 30, no. 7, July 2012, pp. 1193–1204.
[7] R. Myerson, *Game Theory: Analysis of Conflict*, Harvard Univ. Press, 1991.
[8] A. Imran and L. Giupponi, "Use of Learning, Game Theory and Optimization as Biomimetic Approaches for Self-Organization In Heterogeneous Networks," in *Cognitive Signals and Communication Technology*, Springer, 2014, pp. 237–68.
[9] D. Monderer. and L. S. Shapley, "Potential Games," *Games and Economic Behavior*, vol. 14, 1996, pp. 124–43.
[10] Y. Xu *et al.*, "Opportunistic Spectrum Access in Cognitive Radio Networks: Global Optimization Using Local Interaction Games," *IEEE J. Sel. Signal Proc.*, vol. 6, no. 2, Apr. 2012, pp. 180–94.
[11] Y. Xu *et al.*, "Opportunistic Spectrum Access in Unknown Dynamic Environment: A Game-Theoretic Stochastic Learning Solution," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, Apr. 2012, pp. 1380–91.
[12] Z. Du *et al.*, "Exploiting User Demand Diversity in Heterogeneous Wireless Networks," to appear, *IEEE Trans. Wireless Commun.*
[13] M. Bennis and D. Niyato, "A Q-Learning Based Approach to interference Avoidance in Self-Organized Femtocell Networks," *Proc. IEEE GLOBECOM Wksps.*, Dec. 6–10, 2010, pp. 706–10.
[14] Y. Xu *et al.*, "Robust Multiuser Sequential Channel Sensing and Access in Dynamic Cognitive Radio Networks: Potential Games and Stochastic Learning," to appear, *IEEE Trans. Vehicular Tech.*
[15] P. Marshall, *Scalability, Density, and Decision Making in Cognitive Wireless Networks*, Cambridge Univ. Press, 2012.

Biographies

Yuhua Xu (yuhuaenator@gmail.com) received his B.S. degree in communications engineering and Ph.D. degree in communications and information systems from PLA University of Science and Technology in 2006 and 2014, respectively. He is currently with the College of Communications Engineering, PLA University of Science and Technology. His research interests focus on dynamic spectrum access, 5G, game theory, and distributed learning techniques for wireless communications. In 2011 and 2012, he was awarded Certificates of Appreciation as an Exemplary Reviewer for *IEEE Communications Letters*.

Jinlong Wang (wjl543@sina.com) received his B.S. degree in mobile communications, and M.S. and Ph.D. degrees in communications engineering and information systems from the Institute of Communications Engineering, Nanjing, China, in 1983, 1986, and 1992, respectively. He is currently a professor with the College of Communications Engineering, PLA University of Science and Technology. He has widely published in signal processing for wireless communications, information theory, and wireless networks.

Qihui Wu (wuqihui2014@sina.com) received his B.S. degree in communications engineering, and M.S. and Ph.D. degrees in communications and information systems from the Institute of Communications Engineering in 1994, 1997, and 2000, respectively. He is currently a professor at PLA University of Science and Technology, China. His current research interests are algorithms and optimization for cognitive wireless networks, software-defined radio, and wireless communication systems.

Zhiyong Du (duzhiyong2010@gmail.com) received his B.S. degree in electronic information engineering from Wuhan University of Technology, China, in 2009, and his M.S. degree from the Institute of Communications Engineering in 2011. He is currently working toward his Ph.D. degree in communications and information systems at the College of Communications Engineering, PLA University of Science and Technology. His research interests include heterogeneous networks, quality of experience, learning theory, and game theory.

Liang Shen (ShenLiang671104@sina.com) received his B.S. degree in communications engineering and M.S. degree in communications and information system from the Institute of Communications Engineering in 1988 and 1991, respectively. He is currently a professor at PLA University of Science and Technology. His current research interests are information theory, digital signal processing, and wireless networking.

Alagan Anpalagan (alagan@ee.ryerson.ca) received his B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from the University of Toronto, Canada, in 1995,1997, and 2001, respectively. Since August 2001, he has been with Ryerson University, Toronto, Canada, where he co-founded WINCORE laboratory in 2002 and leads the Radio Resource Management and Wireless Access and Networking R&D groups. Currently, he is a full professor and program director for graduate studies in the Department of Electrical and Computer Engineering at Ryerson University.

Cognitive Vehicular Communication for 5G

Shahid Mumtaz, Kazi Mohammed Saidul Huq, Muhammad Ikram Ashraf, Jonathan Rodriguez, Valdemar Monteiro, and Christos Politis

ABSTRACT

Device-to-device (D2D) is increasingly becoming a prominent technology within the 5G story, portrayed as a means of offloading traffic from the core network. The ever increasing demand for vehicular traffic consumption is providing the impetus for a new architectural design that can harness the benefits of D2D for vehicular users, taking a step toward offloading vehicular traffic from the core network. We propose the notion of extending D2D for vehicular scenarios with the potential to coordinate vehicular traffic using the LTE band. Furthermore, we then extend this approach by investigating cognitive radio in synergy with a geo-location database, to exploit white spaces as a means of further offloading vehicular users. Our simulation results have shown that our approach can outperform the legacy IEEE 802.11p in terms of delay.

INTRODUCTION

Each day the world embraces more devices to connect everything, everywhere, and everyone. This kind of interconnecting concedes a huge volume of data traffic among the connected devices. The newly hyped "fifth generation (5G)" paradigm is anticipated to provide the necessary impetus to carry the burden of achieving massive system capacity, reducing latency, and enormously increasing energy saving for the devices.

In addition to the above mentioned expectations, wireless devices in 5G networks are also expected to be constantly interacting with each other as well as with their environment (e.g. data communications from wireless sensors to devices or vice versa). In addition to human-centric D2D communications, one very important use case for D2D is vehicle-to-vehicle (V2V) communications. Recently [1] it has been shown that the integration of information and communication technologies with transportation infrastructure and vehicles will revolutionize the way we travel. Moreover, vehicles are indeed the third place, after homes and offices, where citizens spend the most time daily. V2V communications have already been of the focus of the wireless communications community for many years. For example, IEEE has already developed the 802.11p standard for V2V communication which is based on dedicated short-range communication (DSRC) technology. DSRC technology is mainly used to support intelligent transportation system (ITS) applications in V2V [2] scenarios, but due to the lack of pervasive infrastructure deployment and sufficient transmission range, the IEEE 802.11p standard is normally considered to offer intermittent and short-lived connectivity between vehicles and roadside infrastructure (V2I) [3]. Using DSRC technology to provide V2V communication in a pure distributed fashion may not always guarantee reliability and efficiency in practical applications.

It is commonly accepted that one solution to this problem relies on Long Term Evolution (LTE) technology. However, LTE does not natively sustain V2V communications [4]. For instance, when vehicle density is high, the beaconing signals of vehicular safety applications may easily overload the serving eNB. To handle this issue, a significant amount of such signals should be distributed directly among vehicles, without going through the eNB. In LTE-Advanced (LTE-A), D2D communication is considered to allow direct message delivery between terminals in proximity to lighten the load of eNB [5]. Hence the infrastructure-aided D2D technologies can serve as a natural approach to enable reliable and efficient V2V communications without negatively affecting conventional cellular systems. Current research considers that D2D would be one of the mainstays for 5G networks. To this end, the Third Generation Partnership Project (3GPP) is already in consensus for studying D2D systems rigorously within the LTE Proximity Services study item. This study item belongs to the timeframes between 3GPP Rel-12 and Rel-13.

In general, LTE based cellular systems have no intra-cell interference due to orthogonal subcarriers, but this orthogonality will disappear when D2D users co-exist with other cellular devices. Therefore, there are two approaches to assign radio resource in D2D-based networks. First approach is to assign orthogonal resources

Shahid Mumtaz, Kazi Mohammed Saidul Huq, and Jonathan Rodriguez are with Instituto de Telecomunicações.

Muhammad Ikram Ashraf is with University of Oulu.

Valdemar Monteiro is with Instituto de Telecomunicações and Kingston University.

Christos Politis is with Kingston University.

0163-6804/15/$25.00

We propose a novel cognitive radio based resource allocation policy when applying D2D techniques in V2V. This allocation policy will control the interference between cellular devices and D2D vehicles. Moreover, the decision about the vehicles' communication mode should account for a feasible range under different V2V and eNB distances.

between D2D and other cellular devices (static allocation); second approach is to assign concurrent resources between D2D and other cellular devices (dynamic allocation) [5]. Clearly, the second approach permits a more efficient use of the available radio resources, but it also introduces new interference problems [6].

To this end, we propose a novel cognitive radio-based resource allocation policy when applying D2D techniques in V2V. This allocation policy will control the interference between cellular devices and D2D vehicles. Moreover, the decision about the vehicles' communication mode should account for a feasible range under different V2V and eNB distances. By using D2D it is possible to both reduce the latency and to design a solution that works without cellular network coverage. In D2D mode, vehicles in close proximity communicate directly, which eventually decreases the latency and offloads the traffic from eNBs. D2D will be an appealing solution for local data exchange between vehicles. .

The rest of the article is organized as follows. We discuss why we need LTE-A for vehicular communication, and we present the design aspects for V2V. We then discuss the standardization activities of V2V and the state of the art (SoTA). We then present the system model with simulation results, followed by a discussion of future challenges and the conclusion.

IEEE 802.11P vs LTE-A: Deficiency and Remedies

There are several reasons to choose LTE-A for vehicular communication. The major ones are discussed below [3].

Licensed communication: LTE-A communication is based on the licensed band as compared to IEEE 802.11p, so there will be a control for interference on V2V in LTE-A networks that can easily be manageable, either by the operator or the vehicles.

Coverage: LTE-A relies on the deployment of eNBs, which have coverage of approximately 1000 m, which solves the problem of poor, intermittent, and short-lived connectivity in IEEE 802.11p. For instance, in IEEE 802.11p connection performance suffers severely in non-line of sight (NLOS) environments such as metropolitan city areas where big skyscrapers prevent (shadow or scatter) signals frequently, which brings out fading scenarios. On the contrary, the eNBs in LTE-A networks provide much better performance in NLOS environments due to their position in higher stature.

Scalability: LTE-A networks are accessible for a large number of cellular devices, thanks to its scalable bandwidth, as compared to IEEE 802.11p, which is not scalable for high vehicle density scenarios and also lacks a mechanism to quickly disseminate messages over an increased coverage range.

CAPEX/OPEX: LTE-A uses only one eNB for coverage, which saves CAPEX/OPEX as compared to IEEE 802.11p, which needs many road side units (RSUs) for coverage and to communicating with the Internet.

Capacity: LTE-A offers high downlink and uplink data rates (up to 1 Gps and 500 Mpbs), applying advanced antenna techniques, which eventually supports a higher number of vehicles inside a cell compared to IEEE802.11p, which only supports data rates up to 27 Mb/s.

Infotainment streaming: Future modern vehicles will be capable of exchanging infotainment content (i.e. audio/video streaming, email, software updates) between them, and this will be possible by using D2D communication over the LTE-A band.

Delay: One of the major concerns when considering LTE-A for vehicular communication is delay. LTE-A traffic always crosses infrastructure nodes, even though devices are close to each other. Recent advances in D2D communication in LTE-A mode will solve this problem and offload traffic from infrastructure nodes.

Please see Table 1 [7] for a detailed comparison of LTE-A with other technologies.

D2D Design Aspects for Vehicular Communication

Most of the D2D design aspects described in [5] directly apply to V2V communication in addition to the following enhancements.

The **communication environment** in V2V is quite different than in D2D due to the high mobility of the vehicles. Thus, network connectivity may play a more important role in vehicular communications, compared with system throughput. These characteristics can significantly affect D2D resource allocation strategies and system performance, and thus should be re-examined for V2V.

Scheduling mechanisms envisioned for D2D communications can be used for vehicular communications, but to accommodate these mechanisms in vehicular systems is a not a trivial task. Both uplink and downlink channels must be taken into account while applying D2D scheduling mechanisms to vehicular applications. For the uplink, efficient schedulers must be developed to avoid congestion in crowded networks. For the downlink, a new cross-layer based scheduling is needed in LTE-A to cope with vehicular applications. This can be done by designing a new efficient LTE-A QoS class scheduler [3].

Control plan latency is the time required to perform the transitions between different LTE states. A D2D in LTE is always in one of three states: connected (active), idle, or dormant (battery saving mode). 3GPP specifies that the transition time from the idle state to the connected state should be less than 100 ms, excluding downlink paging and non-access stratum (NAS) signaling delay. Furthermore, it is specified that the transition time from the dormant state to the connected state should take less than 50 ms. Similarly, one way user plan latency in D2D is approximately 5 ms. These latency requirements should be re-redesigned for more strike constraint in the context of vehicular communication where safety applications require every vehicle to transmit a periodic safety message.

Standardization bodies, e.g. ETSI ITS, must rectify their presently available **standards** along

Feature Name	LTE-A	802.11p	Wi-Fi Direct	NFC	ZigBee	Bluetooth	UWB
Standardization	3GPP LTE-A	IEEE	802.11	ISO 13157	802.1504	Bluetooth SIG	802.1503a
Frequency Band	Licensed band	5.86–5.92 GHz	2.4, 5 GHz,	13.56 MHz	868/915 MHz, 2.4 GHz	2.4 GHz	3.1–10.6 GHz
Max transmission distance	1000m	200m	200m	0.2m	10–100m	10–100m	10m
Max data rate	1 Gb/s	27 Mb/s	250 Mb/s	424 kpbs	250 kpbs	24 Mb/s	480 Mb/s
Mobility support	Up to 350 Km/h	Up to 60 Km/h	low	low	low	low	low
QoS	QCI and bearer classes	Enhanced distributed channel access (EDCA)	Enhanced distributed channel access (EDCA)	Enhanced distributed channel access (EDCA)	Enhanced distributed channel access (EDCA)	Enhanced distributed channel access (EDCA)	Enhanced distributed channel access (EDCA)
V2V	**Through D2D**	Ad hoc	Ad hoc	Ad hoc	Ad hoc	Ad hoc	Ad hoc
Vehicle-to-infrastructure (V2I)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Uniformity of service provision	Yes	No	No	No	No	No	No
Application	Offload traffic, public safety, context sharing, local advertising, cellular relay	Context sharing	Context sharing, group gaming, device connection	Contactless payment, Bluetooth and Wi-Fi connections	Home entertainment and control, environmental monitoring	OBject EXchange peripherals connection	Wireless USB, high-definition video, precision location and tracking systems
Infrastructure	Users transfer data directly in licensed band	Users transfer data directly in un-licensed band					
Expenses	CAPEX: no costs as users are using the same terminal. OPEX: very low costs in terms of battery usage.	CAPEX: No costs as users are using the same terminal. OPEX: very low costs in terms of battery usage.					

Table 1. Comparison of various technologies.

with **architectures** to enable D2D to support onboard vehicular applications that provide the impetus for road safety and intelligent vehicular systems.

Economic issues should also be considered when deploying D2D mechanisms onto vehicular applications, because D2D uses licensed spectrum which is not free of charge while exchanging data among the vehicles' owners. Therefore, new business models compatible with market pricing must be envisioned.

V2V Standardization

To achieve V2V safety communications, many consortia involved in industrial, governmental, and university research have created significant opportunities in many projects such as US IntelliDrive, CAMP/VSC-2, CICAS, SafeTrip21, and California PATH [8]. In these projects a category of protocol standards for a special mode of operations in IEEE 802.11 for vehicular networks is designed, called wireless access in vehicular environments (WAVE). These protocols are standardized by IEEE in the IEEE 802.11p and IEEE 1609 protocol set. The IEEE 802.11p [9] is an extension to IEEE 802.11 which includes physical (PHY) and (MAC) layer specifications as well as upper layer protocols for such vehicular networking applications. It inherits simplicity among several characteristics and distributes medium access control mechanisms. Furthermore, these standards are mostly utilized in vehicular on-board units (OBUs) and roadside units (RSUs) such as traffic signals which are normally fixed with transport infrastructure. Apart from the USA, such projects in Japan trying to investigate the deployments aspects of vehicular infrastructure consists of ETC (electronic toll collection) and ongoing rollout for vehicular safety communication. Moreover, such research activities are contributing to ARIB (Association of Radio Industries and Business) and ISO CALM (continuous air-interface long and medium range) standardization. On the other side in the EU, the outcome of such projects is mainly used for standardization activities carried out by industry consortia e.g. C2C-CC (Car 2 Car Communication Consortium) and standardization bodies such as ETSA (European Telecommunication Standards Institute) ITS and ISO (International Organization for Standardization) CALM standardization. In contrast, V2V communications are not natively supported in 3GPP standardization, but given the diverse performance requirements from the wide spectrum of vehicular communication, LTE-A can be an emerging solution for such V2V communication. It has been envisioned to exploit the very existing LTE-A infrastructure to support vehicular networking applications either through advanced LTE-A-enabled OBRs (on-board radios) or using smartphones with LTE-A connectivity. However, the key challenge is to deliver time critical data and efficient resource

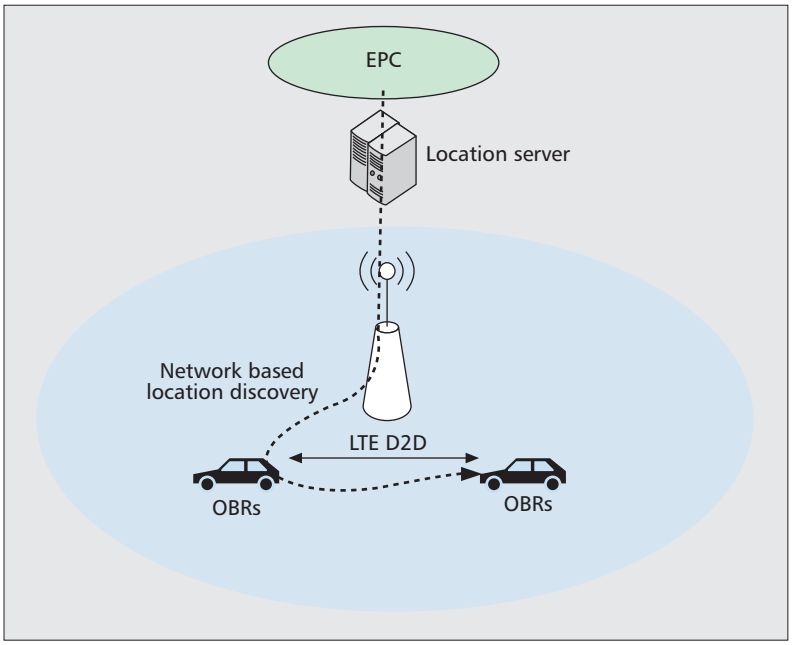


Figure 1. Core architecture V2 V communication.

sharing between vehicles and users over the LTE-A interface.

V2V State of the Art

There have been quite a few works done so far for V2V communication in the LTE-A band, but to the best of our knowledge, no works have been conducted on D2D communication for vehicular technology. D2D communication would be an interesting candidate for local data exchange among vehicles. In [8] the authors provide a detailed review and survey on LTE-A for vehicular networking applications. Different aspects of the suitability of LTE-A over HSUPA cellular networks have been presented in [10]. Similarly, the authors in [11] provide a unified framework to offset the scalability issues by combining LTE-A and cloud-based architecture. In [12] the authors analyze capacity of such LTE-A based vehicular communication networks. Moving on, the authors in [13] devise a software based framework for designing and developing vehicular network applications on smartphones. The authors in [14] anticipated a methodology for 3G cellular network-assisted data delivery for vehicular ad hoc network. The authors in [15] present a heterogeneous network architecture to address smartphone-based information dissemination issues. The authors in [3, 9] present an excellent performance evaluation of IEEE 802.11p and LTE-A.

V2V Functional Block in the LTE-A SAE Architecture

As described in the previous section, there has not been any specific architecture proposed in 3GPP to support V2V communication in LTE-A. In order to do so, we reuse the architecture of D2D communication presented in Fig. 1. In our model, we divide the location services into two parts: location discovery and direct communication. The location discovery mechanism can be network-assisted or vehicle-assisted. It involves the discovery of neighborhood vehicles that are within the LTE-A coverage area. Moreover, location discovery can act as a stand-alone service to vehicles in case of accidents and does not require direct communication. Direct communication enables vehicles to communicate directly without an eNB, within their coverage area. Furthermore, vehicles may initiate direct communication without location discovery. However, location discovery is considered a natural process for direct communication, which reduces the need for manual interaction. Indeed, network supported location discovery is the foremost step for direct communication underlying cellular networks.

In order to assist the existing infrastructure to support V2V communication, it is assumed that vehicles are equipped with advanced LTE-A enabled OBRs. The proposed architecture consists of vehicles, a radio access network, and a core network. To enable the location discovery services, "location server" is embedded into the core network. The location server provides the following functionalities to the proposed V2V based infrastructure:

- Connect between OBRs and the mobile network.
- Identify proximity between vehicles and inform the OBRs about the opportunities.
- V2V session initiation.

The V2V session initiation process is triggered by the location server by sending initiation requests to the MME, while the MME is accountable for the V2V radio bearer setup and delivery of IP addresses of V2V terminating devices, as illustrated in Fig. 2. Moreover, MME provides seamless connectivity operations among higher protocol layers and the mobility process between V2V and cellular networks. Finally, OBRs gather relevant information through periodic exchange of data messages among vehicles controlled by eNBs. However, these data messages are directly communicated over the LTE-A band.

From a data flow perspective, OBRs gather relevant information and periodically exchange date messages with other vehicles via direct communication over the LTE-A band. Please note that vehicles are always under the control of eNBs.

System Description and Model

This section presents the system description, modelling, and sensing algorithm

System Description

LTE-A operates in two modes: a frequency division duplex (FDD) mode and a time division duplex (TDD) mode. The TDD mode supports duplexing UL/DL (uplink/downlink) by allocating time slots in a common band as a function of the service asymmetric level. Therefore, LTE-A TDD is suited to fit the asymmetric services without wasting system capacity. The LTE-A-A FDD mode uses paired uplink and downlink bands and is currently being adopted by European and American operators. In the case of Internet based applications, the traffic patterns

are asymmetric with much lower usage of the uplink band in comparison to the downlink band, which is used for downloading high speed data. This means that LTE-A FDD is a downlink capacity-limited system and consequently UL bands have been underutilized by cellular operators. To confirm this, recent spectrum occupancy measurements performed in Europe pointed out that a power spectrum density (PSD) measured on the UL bands is 20 dB below the DL bands.

System Model

The proposed V2V system exploits the LTE-A-A UL bands; the victim device is the LTE-A eNB, which is likely to be far from the V2V radio (each car has an LTE-A radio) which creates local opportunities due to the transmit power between the V2V transmitter and the LTE-A eNB. These potential opportunities in LTE-A FDD UL bands are in line with the interference temperature metric proposed by the FCC's Spectrum Policy Task Force. The interference temperature model manages interference at the receiver through the interference temperature limit, which is represented by the amount of new interference that the receiver could tolerate. As long as vehicles do not exceed this limit by their transmissions, they can use this spectrum band. However, handling interference is the main challenge in V2V networks, when they are operating on the same band as cellular users. Therefore, the interference temperature concept should be applied in LTE-A-A licensed bands in a very careful manner.

We envisage that a V2V network is able to sense its path loss (i.e. LTE-A radio) between the LTE-A-A eNB and its location. This sensing information is then used by the LTE-A eNB to control the transmit power of a V2V radio in order to avoid harmful interference with the LTE-A-A UL bands. The key issue to enable this is to implement a reliable sensing algorithm and define a strictly non-interference rule for V2V and LTE-A-A coexistence. The resources are centrally controlled by the eNB. In uplink transmission, V2V networks cause interference on the LTE-A eNB if they are operating on the same band or radio resource, as shown in Fig. 3.

In order to avoid harmful interference caused by a V2V network, we propose a simpler approach, i.e. fill part of the available interference temperature with a certain amount of extra interference caused by the D2D network. For simplicity, we consider that the aggregated signals coming from the D2D network are AWGN and cause a noise rise of equal dB, as shown in Fig. 4.

We are considering an LTE-A system operating at FDD 5 MHz of bandwidth, as shown in Fig. 4. An eNB is located in the center of the hexagonal cell. The cellular users act as primary users and the V2V network acts as a secondary user. The LTE-A eNB have an opportunity management entity (OME) which computes the maximum allowable transmit power of each V2V network in order not to disturb the eNB.

The maximum transmit power allowed to a particular V2V network (P_{V2V}) is computed using a non-interference rule that takes into account the aggregated interference of the entire V2V network,

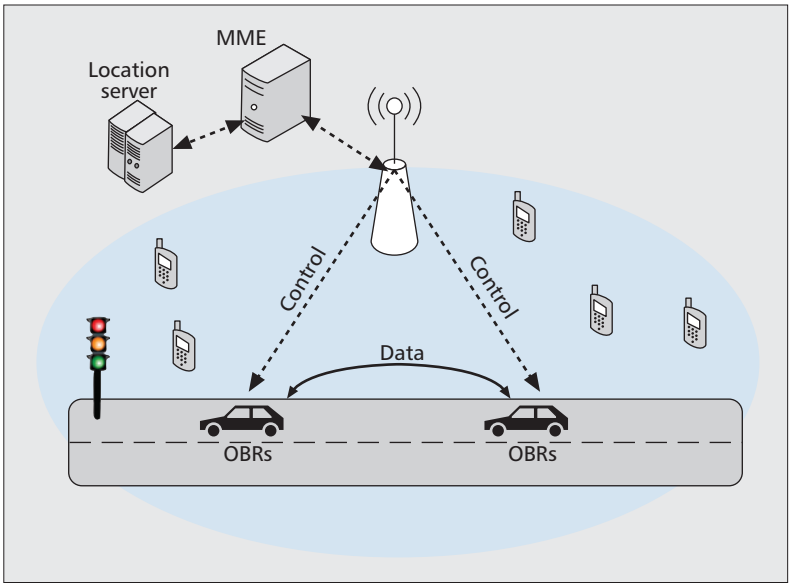


Figure 2. Access architecture for V2V communication.

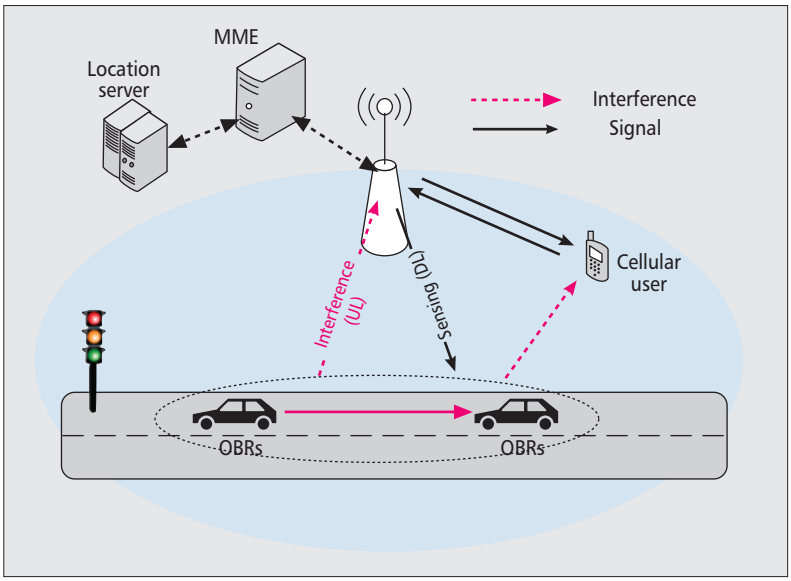


Figure 3. Network scenario.

$$10\log\left(\sum_{k=1}^{K} 10^{\frac{P_{V2V}(k)+G_{V2V}+G_{eNB}-L_p(k)}{10}}\right) \leq 10\log\left(10^{\frac{Nth+\mu}{10}} - 10^{\frac{Nth}{10}} - \Gamma\right) \quad (1)$$

where G_{V2V} and G_{eNB} are the antenna gains of the V2V network and the eNB, respectively, P_{V2V} is the transmit power of the V2V performed by a sensing algorithm, L_p is the estimated path loss between the V2V network and the eNB, K is the number of V2V networks, and Nth is the thermal noise floor. μ is a margin of tolerable extra interference that, by a policy decision, the eNB can bear. Finally, Γ is a safety factor to compensate shadow fading and sensing impairments. Notice if the margin of tolerable interfer-

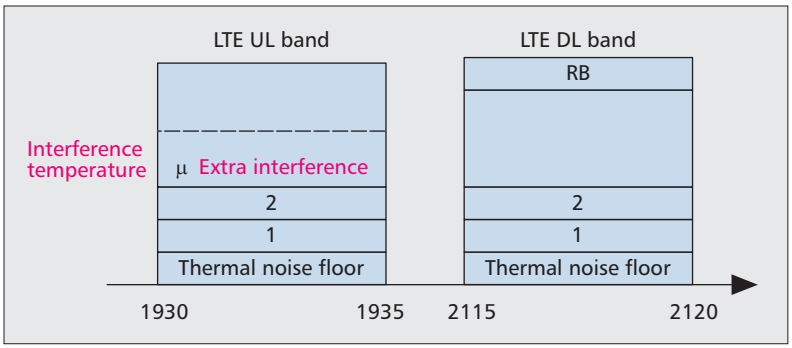


Figure 4. Example of LTE-A-A FDD spectrum band with asymmetric load.

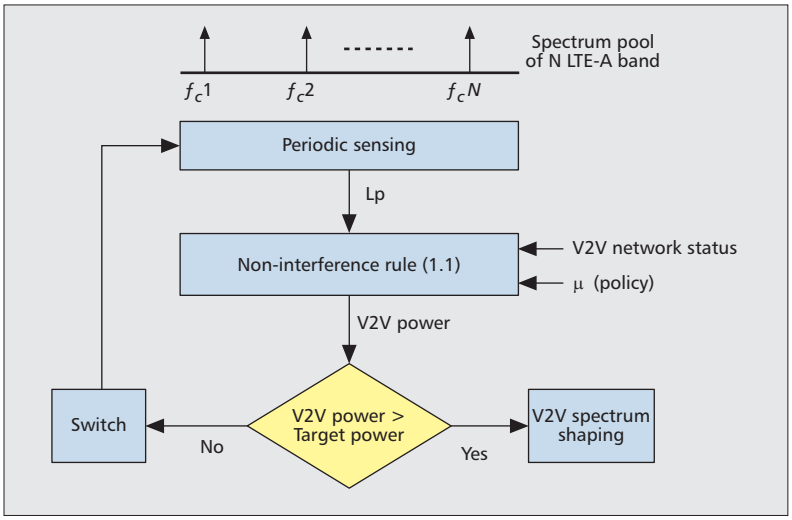


Figure 5. LTE-A spectrum pool mechanism..

ence is set μ = 0, the V2V network must be silent.

It is straightforward to extend this scenario to a 4G multi-operator scenario where several LTE-A UL frequencies cover the same region, as shown in Fig. 5. In this case, the V2V can exploit a spectrum pool of some LTE-A UL carriers. Figure 5 shows a spectrum pool mechanism. As a basic principle, the V2V network is to rent the most appropriate LTE-A UL band to transmit the required power to meet a QoS target. Whenever the required power is sufficient, the V2V signal is formatted by a spectrum shaping module, e.g. using an OFDM modulator. If during the D2D transmission the allowed power becomes lower than the target, the V2V leaves that frequency and switches to another frequency. Sensing is done on a periodic basis to follow the V2V's movement and the correspondent path loss change.

Inter cell interference is computed assuming six adjacent cells, each of them with LTE-A cellular users transmitting its maximum power (21 dBm) in a cell border. The path loss of L_p is calculated as [5] *Path loss* $(L_p) = 40 \log_{10} d + 30 \log_{10} f_c + 49$, where d represents the distance between a sender and a receiver in meters and f_c is the carrier frequency.

Sensing Algorithm

In order to obtain the maximum allowable power for V2V communication, the V2V nodes need to estimate the path loss between the eNB and its particular location. Although we exploit opportunities in the LTE-A UL band, we propose to sense DL signals. This is possible because there is a significant correlation between the average pathloss of the uplink and downlink bands of LTE-A-A. Since the eNB antenna is typically situated in a high location, the DL signal is easier to detect than the multiple UL signals coming from different users. In addition, the DL signal arrives at the sensing antenna in a synchronized manner, which facilitates detection through cyclostationary features of the LTE-A signal. Moreover, sensing and transmission in different bands avoids the allocation of special quiet periods for sensing, as is done in IEEE802.22 systems, booting the D2D spectrum efficiency.

Figure 6 depicts the block diagram of the cyclostationary detector implemented. After a FFT operation, a sliding window of samples performs frequency shifts of $+\alpha/2$ and $-\alpha/2$. The shifted spectrums are then multiplied to obtain the spectrum cyclic density function (SCD). After that, a time smoothing operation is performed through an averaging process during the observation time. The complex values are then squared and integrated over the f domain. Finally, the detection statistic, d, is given by the ratio between the power of the cyclostationary feature, measured at cyclic frequency, α_c, and the estimated noise floor, measured at α_n. In order to estimate this noise floor we take a measurement of the noise at any cyclic frequency, where it is guaranteed there will be no cyclic features present. Notice that as the LTE-A symbol rate is a known cyclic frequency, the algorithm needs to compute only two spectral lines of the SCD functions, α_c and

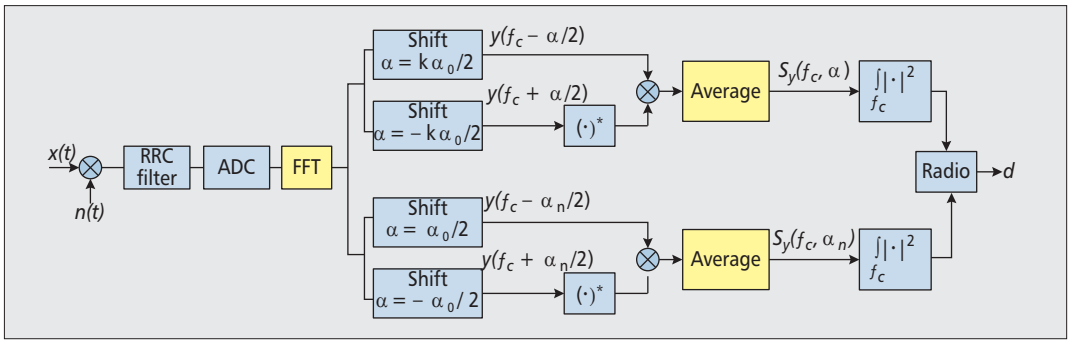


Figure 6. Cyclostationary detector of OFDM signal.

Parameter	Value
eNB transmission power	43 dBm
V2V node transmission power	4 dBm
V2V node speed	50 km/h
Antenna type	Directional (eNB)
Noise power density	–170 dBm/Hz
Noise figure	4 dB
Inter-eNB distance	1000 m
Inter-D2D pair distance	100
V2V pair distance	20 m fixed
Number of cells	7
Carrier frequency	2.6 GHz
System bandwidth	5 MHz
Bandwidth of a channel	180
Number of channels	25

Table 2. Simulation parameter.

α_n, which keeps the detector at a low complexity level.

V2V nodes sense the pathloss between their locations and the eNB. Based on the sensing result and by comparing it with the non-interference rule in Eq. 1, we will categorize the D2D users as interfering and non interfering users.

START:

Step 1: Cognitive Sensing Stage

1) All V2V nodes sense their pathloss from their location to a nearby LTE-A eNB.
2) Compare the transmission power of the V2V nodes with the non-interference rule in Eq. 1.
   a) If P_{V2V} > Eq. 1
   D2D → interfering list
   b) If P_{V2V} < Eq. 1
   V2V → non-interfering list
3) The interfering lists of the V2V nodes are fed back to the eNB.

Step 2: Resource Allocation Stage

1) The LTE-A eNB allocates resource blocks (RBs) to the cellular users and the V2V nodes in the non-interfering V2V list.
2) Based on the unused white space in the LTE-A UL band, the LTE-A eNB allocates available RBs to the V2V nodes in the interfering list.

End.

Figure 7 shows the flow diagram of the sensing algorithm.

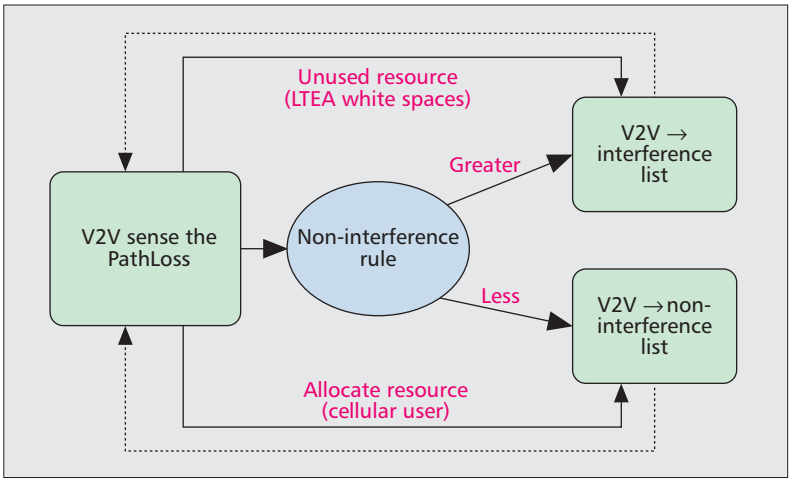


Figure 7. Sensing algorithm flowchart.

Simulation Results and Analysis

We are considering an LTE-A simulation using a system level simulator [7]. Detailed simulation parameters are shown in Table 2. The number of cellular users in each cell is 30, and they are uniformly distributed. There are a total of 20 V2V pairs in each cell, and they are separated from each other at 20 m. Figure 8 shows the E2E delay experienced by the vehicles. It is interesting to know that for a higher number of vehicles, the delay also increases in LTE networks, where traffic goes through eNBs, while in D2D mode the delay stays almost constant as the number of vehicles increases. With these results one can conclude that D2D in LTE mode has good delay constraint for V2V communication.

We also compared (Fig. 9) the proposed resource allocation scheme with a random scheme (i.e. resources are assigned randomly), with the horizontal axis indicating the simulation step in TT1 (i.e. one TT1 in LTE-A is one subframe). As the number of vehicles increases, our proposed scheme has less interference because it uses the unused white space in the UL band.

Research Challenges

There are several challenges and future perspectives that should be considered when designing new efficient V2V communication approaches for 5G, described below.

- 5G networks are expected to contain highly heterogeneous vehicular networks. Therefore, it is important for vehicles to have seamless connectivity across different heterogeneous nodes under time-varying network topology. Hence, next generation vehicles should be more intelligent to support the coexistence of multiple different co-located wireless networks to provide ubiquitous and universal access to broadband services.
- As the volume of V2V communication increases in 5G networks, this will impact the huge data transfer between vehicles and will pose new and unique challenges to data management of vehicular networks.
- Currently GPS is used as a localization system in automobiles. GPS is vulnerable to several

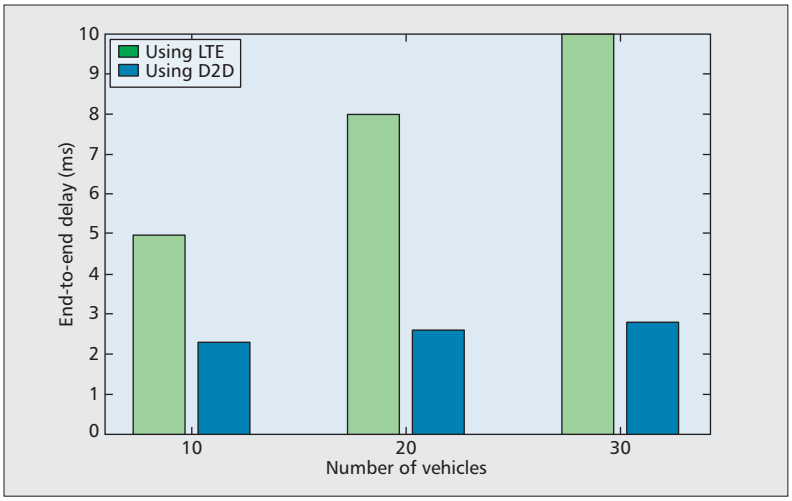


Figure 8. V2V communication delay.

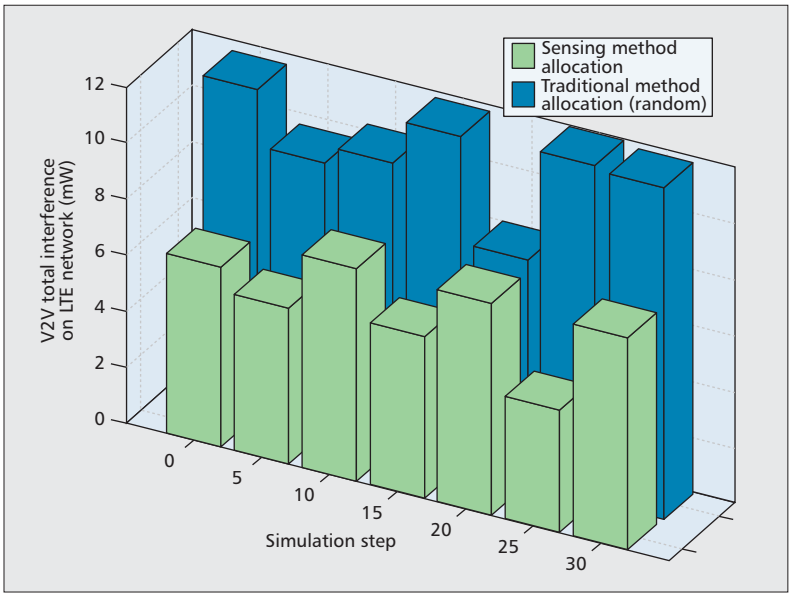


Figure 9. V2V interference analysis.

types of attacks, such as spoofing and blocking. Moreover, GPS signals are unavailable in tunnels and during bad weather. Therefore, there should a new localization system for the vehicles, rather than depending on conventional GPS.

- Due to the greater number of vehicles in future heterogeneous networks, security and privacy of these networks will be a major concern. Therefore, new secure communication protocols must be investigated, taking into consideration the unique characteristics of heterogeneous vehicular networks.
- To decrease the E2E delay in future heterogeneous vehicular networks, D2D should be a part of vehicular communication, i.e. control will be over the licensed band and data between vehicles can be transfered via D2D (either using IEEE 802.11p or the licensed band). This approach increases the coverage and the capacity of future heterogeneous vehicular networks.

Conclusion

In this article we presented the idea of D2D communication for vehicular technology. We showed that D2D communication is a good candidate to decrease the delay constraint in vehicular communication. Then we presented a novel resource allocation scheme that decreases the interference between vehicular nodes and normal cellular users. Finally, we showed by simulation that our scheme outperformed conventional schemes, and that vehicular communication in LTE-D2D mode is a good application for 5G technology. In the future our plan is to provide a comparative study between the proposed schemes and the IEEE802.11p standard.

Acknowledgments:

This work was carried out under the E-COOP project (PEst-OE/EEI/LA0008/2013), funded by national funds through FCT/MEC (PIDDAC) and CARCODE project N. 30345, co-financed by the European Funds (FEDER) by COMPETE, Programa Operacional Factores de Competitividade (POFC) of QREN

References

[1] W. Xing *et al.*, "Resource Allocation Schemes for D2D Communication Used in VANETs," *2014 IEEE 80th Vehic. Tech. Conf. (VTC Fall),* 14–17 Sept. 2014, pp. 1–6.
[2] D. Jiang and L. Delgrossi, "IEEE 802.11p: Towards an International Standard for Wireless Access in Vehicular Environments," *VTC Spring 2008, IEEE Vehic. Tech. Conf., 2008*, 11–14 May 2008, pp. 2036–40.
[3] G. Araniti *et al.*, "LTE for Vehicular Networking: A Survey," *IEEE Commun. Mag.*, vol. 51, no. 5, May 2013, pp. 148–57.
[4] A. Vinel, "3GPP LTE Versus IEEE 802.11p/WAVE: Which Technology is Able to Support Cooperative Vehicular Safety Applications?," *IEEE Wireless Commun. Lett.*, vol. 1, no. 2, Apr. 2012, pp. 125–28.
[5] S. Mumtaz *et al.*, "Direct Mobile-to-Mobile Communication: Paradigm for 5G," *IEEE Wireless Commun.*, vol. 21, no. 5, Oct. 2014, pp. 14–23.
[6] S. Mumtaz *et al.*, "Smart Direct-LTE Communication: An Energy Saving Perspective," *Elsevier Ad Hoc Networks*, vol. 13, Part B, Feb. 2014, pp. 296–311.
[7] D. Feng *et al.*, "Device-to-Device Communications in Cellular Networks," *IEEE Commun. Mag.*, vol. 52, no. 4, Apr. 2014, pp. 49–55.
[8] G. Karagiannis *et al.*, "Vehicular Networking: A Survey and Tutorial on Requirements, Architectures, Challenges, Standards and Solutions," *IEEE Commun. Surveys & Tutorials*, vol. 13, no. 4, 4th Quarter 2011, pp. 584–616.
[9] Z. Mir and F. Filali, "LTE and IEEE 802.11p for Vehicular Networking: A Performance Evaluation," *EURASIP Journal on Wireless Communs. and Networking*, vol. 2014, no. 1, May 2014, p. 89.
[10] H. Kim *et al.*, "A Performance Evaluation of Cellular Network Suitability for VANET," *World Academy of Science, Engineering and Technology, Int'l. Science Index 64*, vol. 6, no. 4, 2012, pp. 1023–26.
[11] S. Kato *et al.*, "Enabling Vehicular Safety Applications over LTE Networks," *2013 Int'l. Conf. Connected Vehicles and Expo (ICCVE)*, 2–6 Dec. 2013, pp. 747–52.
[12] M. Phan, R. Rembarz, and S. Sories, "A Capacity Analysis for the Transmission of Event and Cooperative Awareness Messages in LTE Networks," *Proc. 18th ITS WORLD CONGRESS*, Orlando, FL, 16–20 Oct. 2011.
[13] P. Caballero-Gil, C. Caballero-Gil, and J. Molina-Gil, "Design and Implementation of an Application for Deploying Vehicular Networks with Smartphones," *Int'l. J. Distributed Sensor Networks*, vol. 2013, Dec. 2013, p. e834596.
[14] Q. Zhao *et al.*, "When 3G Meets VANET: 3G-Assisted Data Delivery in VANETs," *IEEE Sensors J.*, vol. 13, no. 10, Oct. 2013, pp. 3575–84.

[15] G. Ferrari *et al.*, "Cross-Network Information Dissemination in VANETs," *2011 11th Int'l. Conf ITS Telecommunications (ITST)*, 23–25 Aug. 2011, pp. 351–56.

Biographies

Shahid Mumtaz [M'13] (dr.shahid.mumtaz@ieee.org) received his M.Sc. degree from the Blekinge Institute of Technology, Sweden, and his Ph.D. degree from the University of Aveiro, Portugal. He is now a senior research engineer at the Instituto de Telecomunicações, Pólo de Aveiro, Portugal, working on EU funded projects. His research interests include MIMO techniques, multi-hop relaying communication, cooperative techniques, cognitive radios, game theory, energy-efficient framework for 4G, position information assisted communication, and joint PHY and MAC layer optimization in the LTE standard. He is the author of several papers published in conferences and journals, and of several books.

Kazi Mohammed Saidul Huq (kazi.saidul@av.it.pt) is a research engineer at the Instituto de Telecomunicações, Pólo de Aveiro, Portugal. He received his bachelor's degree in computer science and engineering from Ahsanullah University of Science & Technology, Bangladesh, in 2003. He obtained his master's and Ph.D. degrees in electrical engineering from Blekinge Institute of Technology, Sweden, in 2006 and the University of Aveiro, Portugal, in 2014, respectively. His research activities include 5G paradigm, backhaul, D2D communication, energy-efficient wireless communication, radio resource management, and MAC layer scheduling. He is the author of several publications including papers in conferences and journals, and book chapters.

Muhammad Ikram Ashraf (ikram@ee.oulu.fi) received his M.Tech degree in telecommunication systems from the University of Oulu, Finland in 2008, and MSc in communication network with distinction (Gold Medal) from Bahria University, Pakistan in 2004. He worked as a research engineer at the Centre for Wireless Communication, Oulu, Finland from 2006 to 2011. From 2011 to 2012 he worked as a senior software engineer at Nokia Oulu, Finland. Currently he is pursuing his Ph.D. degree in communication engineering at the Centre for Wireless Communications, University of Oulu, Finland. His research interests are in the field of heterogeneous networks, D2D communication, radio resource management, social-aware networks, and game theory.

Jonathan Rodriguez [SM'13] (jonathan@av.it.pt) received his master's degree in electronic and electrical engineering and a Ph.D. from the University of Surrey, United Kingdom, in 1998 and 2004, respectively. In 2002 he became a research fellow at the Centre for Communication Systems Research and was responsible for coordinating Surrey's involvement in European research projects under Frameworks 5 and 6. Since 2005 he has been a senior researcher at the Instituto de Telecomunicações, where he founded the 4TELL Wireless Communication Research Group in 2008. He was the project coordinator and technical manager of the FP7 C2POWER and FP7 COGEU projects, respectively, and currently acts as coordinator of several national and international projects. He is the author of more than 200 scientific publications, has served as general chair for several prestigious conferences and workshops, and has carried out consultancy for major manufacturers participating in DVB-T/H and HS-UPA standardization. He is a chartered engineer (IET). His research interests include green communications, network coding, cognitive radio, cooperative networking, radio resource management, and cross-layer design.

Valdemar Monteiro (vmonteiro@av.it.pt) received his degree (five years) and master's degree in electronic and telecommunications from the University of Aveiro (Portugal), in 1999 and 2005 respectively. After his graduation in 2000 he became a research fellow at the Instituto de Telecomunicações–Aveiro, and has worked for international research projects that include IST SAMBA, IST MATRICE, 4MORE, and UNITE. In 2008 he joined CV Movel (Cabo Verde), Cape Verde main mobile operator to work as a switch engineer. He is the author of several papers published in conferences and journals, and has carried out consultancy for operators (Portugal Telecom Inovação) and HSDPA standardization. His research interests include radio access networks for legacy and beyond3G systems with specific emphasis on IP networking, cooperative radio resource management, and PHY/MAC optimization strategies. He is researcher fellow at the Instituto de Telecomunicações and a Ph.D. student at the University of Kingston, UK.

Christos Politis [SM] (c.politis@kingston.ac.uk) is professor (chair) of Wireless Communications at Kingston University London, Faculty of Science, Engineering and Computing (SEC). There he is the co-director (enterprise) of the newly established Digital Information Research Centre (DIRC) with a staff of 25 academics, 20 postdoctoral researchers, and more than 60 Ph.D. students, making it one of the largest in the field in the UK. Upon joining KU as a senior lecturer in 2007, he co-founded and led a research group on wireless multimedia & networking (WMN). He was promoted to reader in 2010 and to full professor in 2015. He teaches modules on wireless communications and networks. Prior to this post, he worked for Ofcom, the UK Regulator and Competition Authority, as a senior research manager. While at the University of Surrey, UK, he undertook a postdoc working on virtual distributed testbeds in the Centre for Communication Systems Research (now the 5G Innovation Centre). This was preceded by placements with Intracom-Telecom SA and Maroussi 2004 SA in Athens, Greece. He has managed to raise several millions of funding from the EU and UK research and technology frameworks under the ICT and Security programs. He holds two patents and has published more than 170 papers in international journals and conferences and chapters in nine books. He sits on the Board of Directors (BoD) of a couple of technology start-ups and advises several governmental and commercial organizations on their research programs/agendas and portfolios. He holds a Ph.D. and MSc. from the University of Surrey, UK, and a B.Eng. from the Technical University of Athens, Greece. He is a senior member of the IEEE, UK chartered engineer, and member of the Technical Chamber of Greece.

CogCell: Cognitive Interplay between 60 GHz Picocells and 2.4/5 GHz Hotspots in the 5G Era

Kishor Chandra, R. Venkatesha Prasad, Bien Quang, and I. G. M. M. Niemegeers

ABSTRACT

The rapid proliferation of wireless communication devices and the emergence of a variety of new applications have triggered investigations into next-generation mobile broadband systems, i.e. 5G. Legacy 2G–4G systems covering large areas were envisioned to serve both indoor and outdoor environments. However, in the 5G era, 80 percent of all traffic is expected to be generated indoors. Hence, the current approach of macrocell mobile networks, where there is no differentiation between indoors and outdoors, needs to be reconsidered. We envision 60 GHz mmWave picocell architecture to support high-speed indoor and hotspot communications. We envisage the 5G indoor network as a combination of, and interplay between, 2.4/5 GHz having robust coverage and 60 GHz links offering a high data rate. This requires intelligent coordination and cooperation. We propose a 60 GHz picocellular network architecture, called CogCell, leveraging ubiquitous WiFi. We propose to use 60 GHz for the data plane and 2.4/5GHz for the control plane. The hybrid network architecture considers an opportunistic fall-back to 2.4/5 GHz in case of poor connectivity in the 60 GHz domain. Further, to avoid the frequent re-beamforming in 60 GHz directional links due to mobility, we propose a cognitive module, a sensor-assisted intelligent beam switching procedure, that reduces communication overhead. We believe that the CogCell concept will help future indoor communications and possibly outdoor hotspots, where mobile stations and access points collaborate with each other to improve the user experience.

INTRODUCTION

The unprecedented but anticipated massive growth of mobile data traffic is posing many challenges for 5G communication systems. 5G networks aim to achieve ubiquitous communication between anybody and anything, anywhere and at anytime. The performance requirements are far beyond what is offered by current systems; in particular, a 1000x increase in network capacity is targeted. All this requires new network architecture and technologies. Moreover, new spectrum will be needed. For example, millimeter wave (mmWave) communication requires very different approaches for the PHY, MAC, and network layers. The general consensus among researchers and industry is that 5G will not be a mere incremental evolution of 4G [1]. However, 2G–4G will have to be integrated with the new technologies to ensure the support of legacy systems.

Figure 1 shows the 5G communication scenario, where multiple radio access technologies (RATs), i.e. 60 GHz wireless local area networks (WLANs), 2.4/5 GHz WiFi, 28–30 or 38–40 GHz outdoor mmWave base stations (BSs), and macrocell femtocell BSs, are present. For efficient spectrum utilization, multiple licensed as well as unlicensed bands will need to work in cohesion for different applications. mmWave based mobile communication (28–32 GHz and 38–42 GHz spectrum) and WLANs at 60 GHz will coexist with legacy cellular networks and WLANs. Thus, 5G spectrum would span from sub-GHz to mmWave frequency bands to support diverse applications and services. To exploit the available spectrum across the various frequency bands, a highly flexible communication interface is required that can support multiple RATs for various, possibly very different, services at the same time. To meet the above stated requirements, various solutions are being discussed. We summarize them as follows.

Network architecture: Instead of a rigid and infrastructure-centric approach adopted by previous generations, device-centric and user-centric architectures are being advocated for 5G, in order to better support ubiquitous and seamless communication. Further, the concept of cloud-based radio access networks (C-RANs) is proposed to reduce operational costs by efficient utilization of radio resources [2]. In C-RAN, traditional base station functionality such as baseband processing and resource allocation is

The authors are with Delft University of Technology, The Netherlands.

0163-6804/15/$25.00 © 2015 IEEE

offloaded to a central location, to provide dynamic resource allocation leading to a better utilization of baseband processing resources. Another architectural change expected is the macro-assisted small cells, also called *phantom cells* [3]. In this approach, the control plane and data plane are decoupled. The macrocell covering a large area is responsible for the control and management functions, while small cells are used solely for providing high data rate communications. Usually small cells remain in a turn-off state to save energy. Furthermore, for devices that are in proximity of each other, direct device-to-device (D2D) communication is considered and is expected to become an integral part of 5G.

Medium access control and signaling: 5G needs to support a variety of applications that are very different in terms of traffic patterns, data rates, and latency constraints. For example, machine-to-machine (M2M) communication will have infrequent small packets with low data rates but with critical latency requirements. Video applications, e.g. 4K video, have some latency requirements, can tolerate errors to an extent, but will require very high data rates. Web browsing and file sharing applications, on the other hand, have different requirements. In the case of M2M, signaling and control mechanisms employed in current networks would cause high overheads. Widespread use of M2M may lead to situations where thousands of devices try to access a channel simultaneously. Current access mechanisms are not designed for this. Furthermore, to enable D2D, very efficient signaling mechanisms are required so spectrum utilization can be increased. In the case of ultra-dense networks, coordination among small cells, needed to mitigate interference, will lead to high signaling overhead. Thus, flexible medium access and signaling protocols are needed to optimize channel utilization for a wide variety of applications.

Physical layer techniques: From the perspective of the physical layer, to combat the scarcity of available radio spectrum in the lower frequency bands, mmWave frequencies (30 GHz to 300 GHz) are being explored as alternatives for both outdoor and indoor communication due to the huge bandwidth they provide. Licensed 28-30 GHz and 38–42 GHz bands are suitable for outdoor cellular networks [4], while the unlicensed 60 GHz band is suitable for indoor communication due to its propagation characteristics [5]. Another breakthrough technology that will certainly have a distinct place in 5G is massive MIMO [6, 7]. In massive MIMO the number of antennas at a BS is much higher than the number of devices being served, enabling simple *spatial multiplexing* and *demultiplexing*. The small size of antennas and antenna spacing at mmWave frequencies make massive MIMO a suitable beamforming technology for devices[1] as well as BSs.

It is predicted that by the year 2020 indoor/hotspot traffic will account for 80 to 90 percent of total traffic volume [8]. Data rates on the order of multi-Gb/s will be required in indoor environments to support high definition video streaming and gaming applications. Existing 3G and 4G systems were designed to support the same set of services both in indoor and outdoor environments. However, this will not be the case in 5G. A variety of services are emerging and many of them, in particular high data rate uncompressed video, will be mainly confined to indoors and hotspots. Therefore, 5G networks must take care of the traffic dissimilarities between indoor and outdoor environments. To tackle this challenge, high capacity indoor local small cells need to be designed that can provide multi-Gb/s connectivity with better coverage.

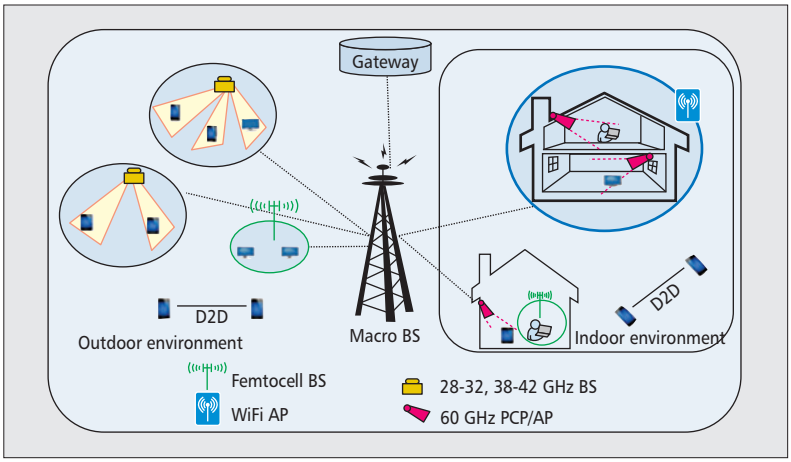


Figure 1. A 5G scenario with multiple radio access technologies.

The 60 GHz frequency band has emerged as the most promising candidate for high speed indoor communications. However, its inability to penetrate walls poses a serious challenge for providing seamless connectivity. Further, the use of narrow beamforming makes it challenging to support mobile devices, due to the link outages caused by antenna beam misalignment resulting from the mobility of users. This requires beam tracking and adaptive beamforming. We propose the CogCell concept, a 2.4 GHz assisted 60 GHz picocellular network architecture in which 60 GHz is used for high speed data communication (data plane traffic) while 2.4/5 GHz WiFi is used for control purposes (control plane traffic). Several 60 GHz picocells are managed by a single WiFi cell, thus facilitating easy and robust network and mobility management with picocells. In the absence of a 60 GHz link, 2.4 GHz can also be used as a fall-back data plane option in CogCell, leveraging the best of both worlds. The problem of frequent re-beamforming in 60 GHz can be circumvented by leveraging the sensing and processing capabilities of smart devices that are using the 60 GHz links. We will show how motion sensors (present in smart phones and tablets) will be used to predict user movement and thus maintain beam alignment. The CogCell architecture has many features, including:

- Better spectrum utilization by switching between the 2.4 GHz and 60 GHz bands for control and data transmission, respectively.
- Opportunistic fall-back to the 2.4 GHz band for data transmission, if the 60 GHz link is not available.
- Sensor assisted cognitive and adaptive beam-tracking, which reduces the need for frequent re-beamforming of 60 GHz links in case user devices move.

[1] We use the term device to mean a mobile station or a handheld user equipment.

The 60 GHz MAC standards IEEE 802.15.3c and IEEE 802.11ad have already been completed, providing data rates up to 5 Gb/s to 7 Gb/s for a range of 10 m to 20 m. IEEE 802.11ad is backward compatible with IEEE 802.11b/g/n/ac. However, there remain several issues that need to be addressed to realize multi-Gb/s 60 GHz indoor networks.

60 GHz Communication for Multi-Gb/s Indoor Connectivity

Despite very sophisticated PHY/MAC layer techniques such as MU-MIMO, higher order modulations, channel bonding, and frame aggregation, it is hard to improve the WiFi data rate further. For example, despite using channel bonding and multi-user MIMO schemes, IEEE 802.11ac can only provide a peak data rate of around 1 Gb/s because of limited bandwidth in the 2.4/ 5 GHz frequency bands. On the other hand, large bandwidth is available in the unlicensed 60 GHz band. The 60 GHz MAC standards IEEE 802.15.3c [9] and IEEE 802.11ad [10] have already been completed, providing data rates up to 5 Gb/s to 7 Gb/s for a range of 10 m to 20 m. IEEE 802.11ad is backward compatible with IEEE 802.11b/g/n/ac. However, there remain several issues that need to be addressed to realize multi-Gb/s 60 GHz indoor networks.

Access delay: 60 GHz devices and access points (AP) employ directional antennas to compensate for free space path loss. IEEE 802.11ad and IEEE 802.15.3c divide the area around an AP into sectors, e.g. a sector can span over 60° or 90°. CSMA/CA based random access is used during predefined time periods — in each sector in a round robin fashion — called contention based access periods (CBAPs). A device has to wait for the CBAP period allocated to its sector. For example, if each sector spans an angle of 90° then there are four sectors. Thus, if a device generates a request just after the allocated CBAP period for its sector, it has to wait until the next three CBAP periods. This could introduce a considerable amount of delay before the request is fulfilled.

Re-beamforming: Although the peak PHY data rate promised by IEEE 802.11ad is about 7 Gb/s, realizing a seamless multi-Gb/s WLAN system providing a sustained peak data rate is difficult. 60 GHz links are highly susceptible to blockage caused by obstacles such as humans, furniture, walls, etc. Further, communication using narrow beams has to track moving devices to maintain the link. With narrow beams, beam misalignment caused by small movements may result in broken links. If a device moves away from the beam coverage area, an exhaustive beam-search is required, resulting in excessive delays and communication overhead. Therefore it is important to keep beam alignment in order to maintain a stable link.

Hand-off: While using directional antennas at 60 GHz, AP/device discovery and fast handover are difficult. Since 60 GHz signals cannot penetrate walls, there will be many 60 GHz APs in an indoor area. This can result in frequent hand-off when a user moves in the indoor area. When moving from one room to another, one should be able to quickly reconnect with another AP. To ensure this, fast discovery and authentication are needed. Since the data rate is very high, a small interruption in signal coverage can lead to the loss of a large amount of data. Further, frequent device discovery and association could lead to excessive energy consumption, resulting in fast battery drain.

To address the above issues, we propose to use a WiFi and mmWave CogCell hybrid architecture. This will enable smooth network management and fast channel access and device discovery. Here WiFi supports control plane functions while 60 GHz offers data plane functionality. To avoid frequent re-beamforming caused by mobility, we employ motion sensors to predict the next location of the user so that appropriate beam switching can be performed.

Indoor Networks Based on a Combination of WiFi and 60 GHz Communication

In this section we discuss the capacity and coverage limitations of 2.4/5 GHz and 60 GHz signals, respectively. We illustrate that 2.4 GHz and 60 GHz systems are complementary in terms of coverage and capacity, and explain how the proposed CogCell architecture enables the interplay of both to provide robust multi-Gb/s WLAN connectivity.

Complementarity of 2.4 GHz and 60 GHz

Figure 2a shows the coverage of 2.4 GHz (left) and 60 GHz (right) signals in an indoor environment. A radio-wave propagation simulator (RPS) [11] employing ray tracing is used to determine the coverage in the indoor area. To calculate the signal power, reflections, up to second order, are considered and all the antennas are assumed to be omnidirectional. The transmission power of antennas is 10 dBm. It is clear that three antennas operating at 2.4 GHz are sufficient to cover the whole area. On the other hand, at 60 GHz every room needs a dedicated 60 GHz antenna, because signal propagation characteristics are significantly different at 2.4 GHz and 60 GHz. mmWaves at 60 GHz do not penetrate through walls. A significant fraction of signal power is absorbed by the walls. This is illustrated by the black ellipses over the blue colored areas in Fig. 2a.

Figure 2b compares the maximum data rates promised by different WLAN standards operating at 2.4/5 GHz and 60 GHz frequency bands. Even though IEEE 802.11n and IEEE 802.11ac use very sophisticated PHY layer techniques such as MIMO, MU-MIMO, channel bonding, and frame aggregation at the MAC layer, the expected data rate is much lower compared to what can be achieved at the 60 GHz frequency band.

It is evident from Fig. 2 that the 2.4 GHz and 60 GHz signals complement each other in terms of capacity and coverage. The capacity of 60 GHz signals is at least 10 times higher than the 2.4 GHz systems. Thus, a hybrid solution, involving 2.4 GHz transmission assisting the 60 GHz devices, can be very effective. Almost every consumer electronic device, such as smartphones, tablets, laptops, cameras, etc., is equipped with WiFi and this trend is expected to continue. Hence, assistance of the 2.4/5 GHz band for 60 GHz communications seems a pragmatic solution.

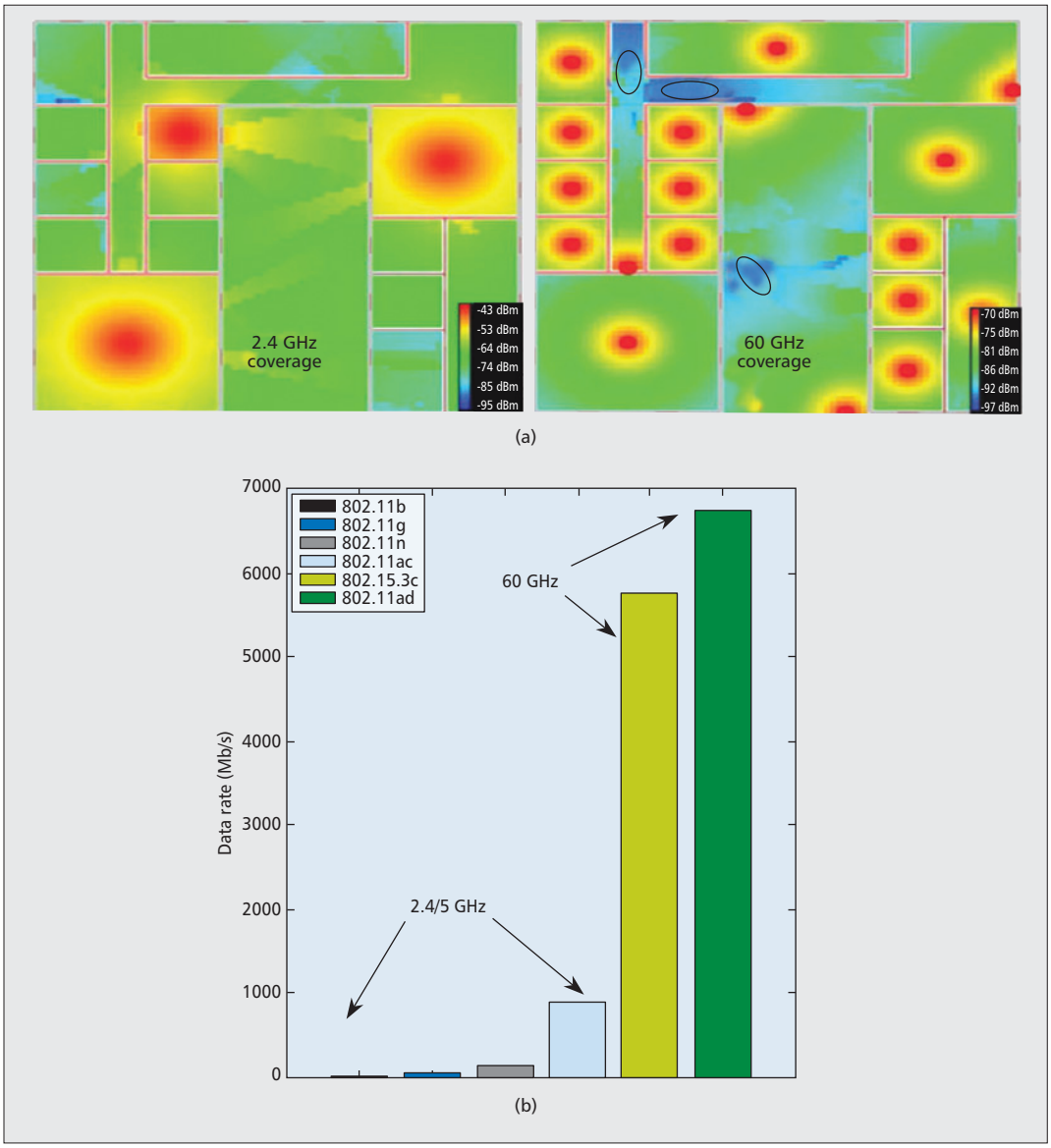


Figure 2. Comparison of signal coverage and offered datarates at 2.4/5 GHz and 60 GHz: a) signal coverage comparison at 2.4 GHz and 60 GHz; b) comparison of peak datarates at 2.4/5 GHz and 60 GHz.

Almost every consumer electronic device, such as smartphones, tablets, laptops, cameras, etc., is equipped with WiFi and this trend is expected to continue. Hence, assistance of the 2.4/5 GHz band for 60 GHz communications seems a pragmatic solution.

Hybrid 2.4 GHz and 60 GHz WLAN Architecture

There can be two types of solutions:

- Utilizing the existing 2.4 GHz WiFi and IEEE 802.11ad, and modify them accordingly; or
- A new system other than IEEE 802.11b/g/n and IEEE 802.11ad. The former category is more likely to succeed as a majority of wireless communication devices are already equipped with IEEE 802.11b/g/n.

One possible approach in the first category could be to use WiFi as a supportive technology to manage the 60 GHz network. The WiFi AP can cover several 60 GHz APs and hence, several 60 GHz APs can be managed by a single WiFi AP. This is the basic idea behind the proposed CogCell architecture. We propose to split the control plane and data plane over 2.4 GHz and 60 GHz, respectively. This scheme is similar to the concept of *phantom cells* [3] proposed for 5G networks.

Figure 3a shows the conceptual diagram of the CogCell architecture. One 2.4 GHz AP covers all the rooms. Further, every room has a 60 GHz PCP/AP (802.11ad APs are called PCP/APs) dedicated for high speed data transmission. In a smaller indoor area such as small homes, a single 2.4/5 GHz AP can be sufficient to provide the coverage, but if the indoor area is large (e.g. a large office, shopping malls, or airports), multiple 2.4/5 GHz APs would be needed to cover the complete area. Moreover, when areas are separated by walls, they always require separate 60 GHz PCP/APs.

In the proposed CogCell architecture, device discovery, association, and channel access requests are transmitted over the 2.4 GHz channel, while data is transmitted over the 60 GHz channel. If a device wants to transmit data, it first sends its request using the 2.4 GHz frequency band. Thereafter, the appropriate 60 GHz AP is directed to facilitate the high speed data transmission. IEEE 802.11ad PCP/APs are tri-band

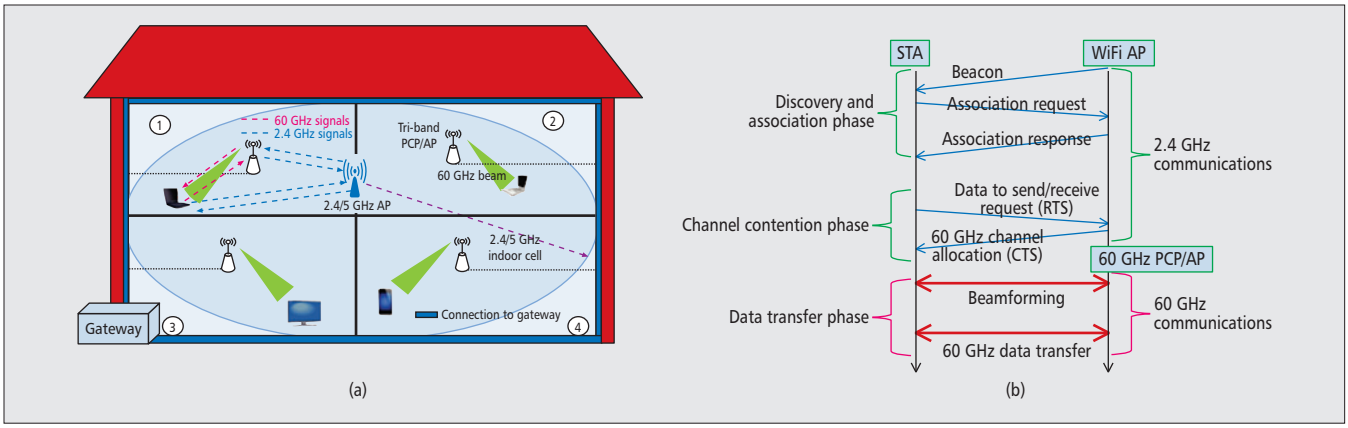


Figure 3. Interplay of 2.4 GHz and 60 GHz frequency bands in the proposed indoor network in 5G: a) network architecture; and b) sequence diagram of dual band transmission.

devices, hence a WiFi AP can communicate with a 60 GHz PCP/AP over 2.4/5 GHz. Figure 3 shows the schematic of signal transmission in the CogCell architecture.

Advantages and Challenges of WiFi and 60 GHz Interplay

It is to be noted that, other than WiFi, LTE may also assist the mmWave communications (LTE-WiGig) [12]. Especially outdoors, LTE can provide better control functionality instead of WiFi due to its limited range. However, in indoor environments, exploitation of WiFi would be more suitable instead of LTE due to the prevalence of WiFi networks over licensed LTE cells. Furthermore, WiFi would be more suitable for indoor mobility management due to its localization capabilities, which are accurate up to a meter and can help in handover between 60 GHz APs, where room level positioning accuracies would be required. When LTE is used in conjunction with WiGig, the data path could be via LTE base stations, or there must be different backhaul connectivity to the WiGig AP. In the first case, an LTE BS would be the bottleneck and it would defeat the purpose of having WiGig. In the latter case, where a backhaul is used for data path via a WiGig AP, then it would indeed be similar to CogCell except that LTE handles the control (rather than a WiFi AP as in the CogCell). LTE-WiGig, of course, helps in outdoor environments and it can provide a high data rate if backhaul connectivity exists. Now we briefly describe the advantages and challenges of the interplay between WiFi and 60 GHz.

Advantages: There are many advantages in a hybrid 2.4 GHz and 60 GHz WLAN system. First, isolated (behind the walls) 60 GHz APs can still facilitate a seamless WLAN experience to indoor users. Second, device discovery and association can be easily performed over 2.4 GHz. As users move from one room to another room, they are still under the same 2.4 GHz APs. Third, information sent over the 2.4 GHz channel can also help in the 60 GHz beamforming procedure. Instead of using two-level exhaustive beam searching, as in IEEE 802.11ad, devices can estimate the approximate direction of each other using 2.4 GHz frames.

Generally, 2.4/5 GHz communications (IEEE 802.11n, IEEE 802.11ac) employ multiple antennas in which approximate direction of arrival can be obtained. Using this rough estimate of direction of arrival, the search space of exhaustive beam searching for 60 GHz is reduced. A similar approach has been employed in [13], which shows that inferring the direction of 60 GHz transmission using 2.4/5 GHz can reduce the link setup overhead by avoiding exhaustive beam searching.

Figure 4a shows results from MATLAB simulations for the WiFi assisted device discovery mechanism, assuming devices can infer rough sector estimates using 2.4 GHz transmissions. The beamwidth of all the devices and PCP/APs is assumed to be 60°. All the parameters are listed in Table 1. The results are compared with the stand-alone 60 GHz directional device discovery scheme proposed in [14]. It can be observed that the WiFi assisted scheme is nearly 150 percent to 300 percent faster than the 60 GHz directional device discovery scheme. The results also show the effect of signaling overhead due to 2.4 GHz control frame transmission, which is obtained by including the time required for transmission of extra management frames over 2.4 GHz.

Furthermore, the CogCell architecture can reduce channel access delay because a device can place the data transmission request over the 2.4 GHz channel whenever it wants. On the other hand, in sectorized MAC protocols such as IEEE 802.11ad, a device has to wait for channel access if the 60 GHz AP is serving a different sector.

Challenges: The hybrid 2.4/5 GHz and 60 GHz network also poses many challenges. First, an increased number of WiFi devices can hinder control plane communication. To address this issue, we propose to prioritize the 60 GHz channel access requests over the 2.4 GHz requests. We define two categories of frames sent over the 2.4 GHz channel:

- 60 GHz channel request frames.
- 2.4 GHz channel request frames by non-60 GHz devices.

We assign different contention window sizes and allowed maximum number of retransmissions for these categories, which are shown in Table 1.

Figure 4b and Fig. 4c show the MATLAB simulation results for the average channel access delays and transmission probabilities for both types of requests. It can be seen that a significantly faster channel access and higher transmission probabilities can be guaranteed for the 60 GHz channel requests.

Second, power consumption of multiple radios working simultaneously can drain the batteries of mobile devices. Hence, novel schemes are required to reduce device power consumption. One possible solution could be to turn on the 60 GHz radio only when data plane communication is required. Third, the 2.4/5 GHz control plane would also be used as a fall-back option if the 60 GHz data plane is not available. This requires an intelligent mechanism to determine when the data plane fall-back should be triggered as 60 GHz link quality can deteriorate due to multiple reasons such as antenna misalignment due to user movement, blockage due to obstacles, etc.

Sensor-Assisted Intelligent Beam Switching

Communication using narrow beam directional antennas can cause frequent link degradation due to device movement. This is particularly the case with handheld devices such as smartphones, tablets, etc. To set up the directional link between two devices, IEEE 802.11ad provides a beamforming mechanism for the selection of the best transmit and receive antenna-beam pair. In the case of device mobility, beam alignment can be disturbed, whichs could result in frequent outages of links. If the link quality degrades below a certain limit, the mechanism to select the best beam-pair is restarted (we call this re-beamforming). The re-beamforming procedure involves exhaustive searching in all the possible transmit and receive directions. This leads to a considerable amount of communication overhead as well as degradation of quality of service (QoS).

If the next position of the users is known, the PCP/AP and the device can switch their beams to the appropriate beam sectors. We propose to use the motion sensors, such as the accelerometer and gyroscope, to identify device movements and predict the next location of the device. These sensors are already embedded in most modern devices, hence this method is not unrealistic and is economically viable. To retrieve the useful information from these sensors it is possible to combine the data from two or more sensors. Such combination of sensors is referred to as a virtual sensor. The *rotation vector sensor* is such a virtual sensor, where the accelerometer, gyroscope, and magnetometer data are fused. The rotation vector sensor gives the orientation of the device relative to the East-North-up coordinates. The azimuth angle from this sensor can be used as an indication of the direction of the user, which can assist in identifying the next beam-pairs.

Figure 5a shows the system diagram of sensor assisted beamforming. Whenever a movement occurs, based on the gathered sensor data, the next location of the user is predicted and beam

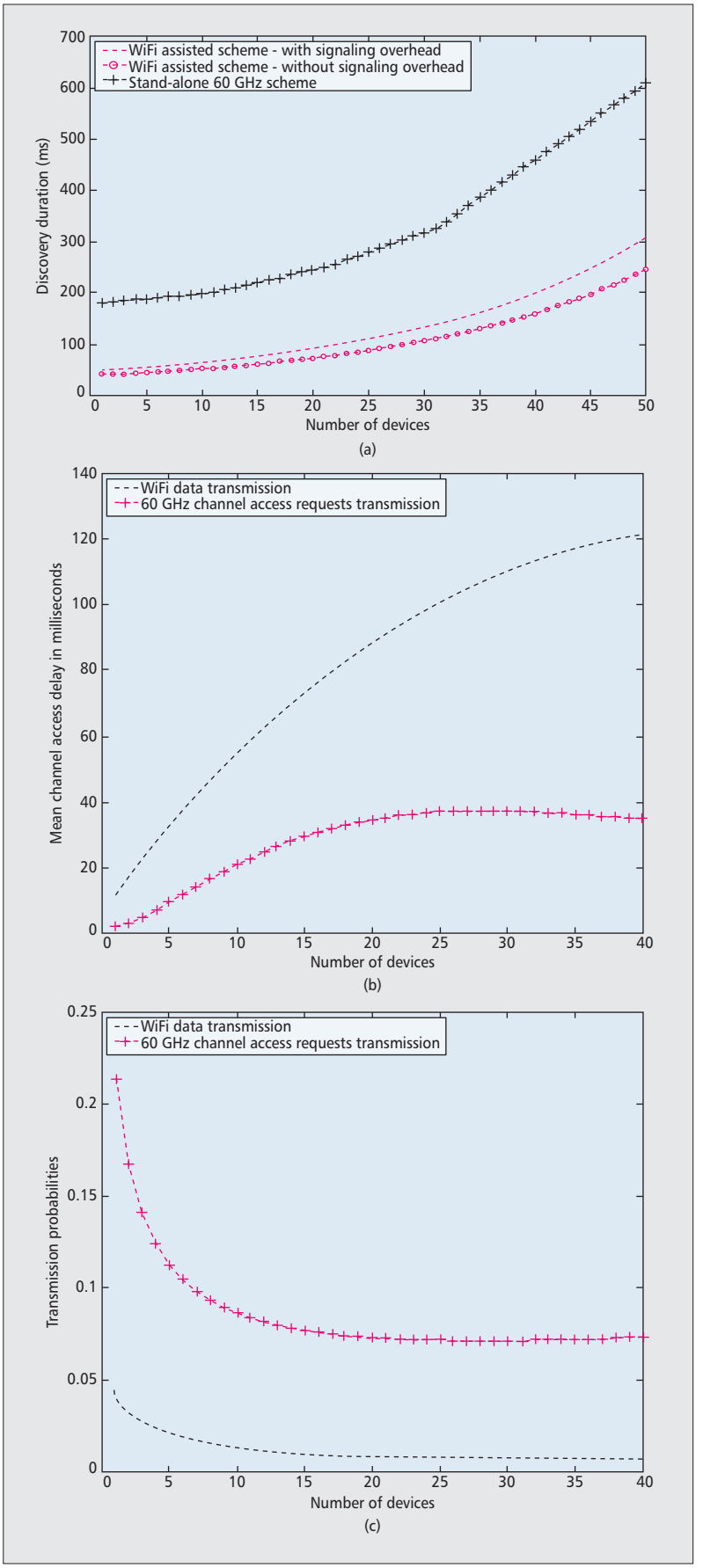


Figure 4. Device discovery time comparison and channel access performance: a) average device discovery vs number of devices; b) channel access delay; and c) transmission probabilities.

We believe that the combination of 2.4/5 GHz WiFi and 60 GHz communication will play an important role in indoor networks in the 5G era, and we showed the approach to exploit them together.

Parameters	Typical values	Parameters	Typical values
Control frame transmission rate	1 Mb/s	Retry limit [2.4 GHz]	5
WiFi data rate	54 Mb/s	CW_{max} [60 GHz]	16
SIFS [2.4 GHz]	10μs	CW_{max} [2.4 GHz]	256
SIFS[60 GHz]	3μs	RTS	20 Bytes
Slot time [2.4 GHz]	20μs	CTS	14 Bytes
Slot time [60 GHz]	5μs	ACK	14 Bytes
DIFS [60 GHz]	SIFS + slot time	PHY header	16 Bytes
DIFS [2.4 GHz]	SIFS + 2 × slot time	MAC header	24 Bytes
RIFS	300 μs	WiFi data	1024 Bytes
CW_{min} [60 GHz]	8	Association request	1024 Bytes
CW_{min} [2.4 GHz]	32	Association response	16 Bytes
Retry limit [60 GHz]	5	Sector sweep and feedback frame	1024 Bytes

Table 1. MAC parameters for prioritized control channel access.

switching is performed to maintain beam alignment. Figure 5a shows the preliminary simulation results when a device moves along the stated route in [15]. When the PCP/AP beamwidth is 30°, 14 instances of re-beamforming are required without using sensor data. On the other hand, with the help of sensor prediction, the number of re-beamformings can be reduced to four. Similarly, when the PCP/AP beamwidth is 20°, instead of 18 re-beamforming instances without sensor prediction, re-beamforming is needed only five times using the rotation vector sensor data. In this simulation, we assumed that the PCP/AP knows the sensor information. However, in practical scenarios, sensor information needs to be communicated to the PCP/AP. This can be done by including sensor information in the 802.11ad data frames. This preliminary examination of using sensor data for beam switching seems encouraging and requires further investigation.

Conclusions

In this article we proposed a novel indoor network architecture, CogCell, for 5G. The proposed CogCell architecture enables the interplay between 2.4 GHz and 60 GHz bands for control plane and data plane transmissions, respectively. CogCell promises a robust multi-Gb/s WLAN experience at 60 GHz frequency bands enabling faster device discovery and medium access. We believe that the combination of 2.4/5 GHz WiFi and 60 GHz communication will play an important role in indoor networks in the 5G era, and we showed the approach to exploit them together. Further, a sensor-assisted intelligent beam switching scheme for 60 GHz communication was proposed. It was shown that with the help of rotation-vector sensor-data, frequent re-beamforming in the 60 GHz directional links can be significantly reduced. Thus, link maintainability in 60 GHz is guaranteed, resulting in fewer requests on WiFi APs leading to efficient use of 60 GHz and WiFi.

References

[1] J. Andrews *et al.*, "What 5G Will Be?," *IEEE JSAC*, Sept. 2014.
[2] C.-L. I *et al.*, "Toward Green and Soft: A 5G Perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, 2014, pp. 66–73.
[3] Y. Kishiyama *et al.*, "Future Steps of LTE-A: Evolution Toward Integration of Local Area and Wide Area Systems," *IEEE Wireless Commun.*, vol. 20, no. 1, 2013, pp. 12–18.
[4] T. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, 2013, pp. 335–49.
[5] T. Rappaport, J. Murdock, and F. Gutierrez, "State of the Art in 60-GHz Integrated Circuits and Systems for Wireless Communications," *Proc. IEEE*, vol. 99, no. 8, Aug 2011, pp. 1390–436.
[6] W. Roh *et al.*, "Millimeter-Wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results," *IEEE Commun. Mag., IEEE*, vol. 52, no. 2, pp. 106–113, Feb. 2014.
[7] F. Rusek *et al.*, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," *CoRR*, vol. abs/1201.3210, 2012.
[8] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2018," Feb. 2014.
[9] "IEEE 802.15.3c working group, tgc3,," Report.
[10] "Draft Standard — Part 11:Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY)specifications — Amendment 4: Enhancements for Very High Throughput in the 60GHz Band," IEEE P802.11adTM/D9.0, July 2012.
[11] J. Deissne *et al.*, "RPS Radiowave Propagation Simulator User Manual-Version 5.4," Actix GmbH, 2008.
[12] Available: http://newsroom.intel.com/community/intel newsroom/blog/2015/03/02/intel-launches-new-mobile-socs-lte-solution.

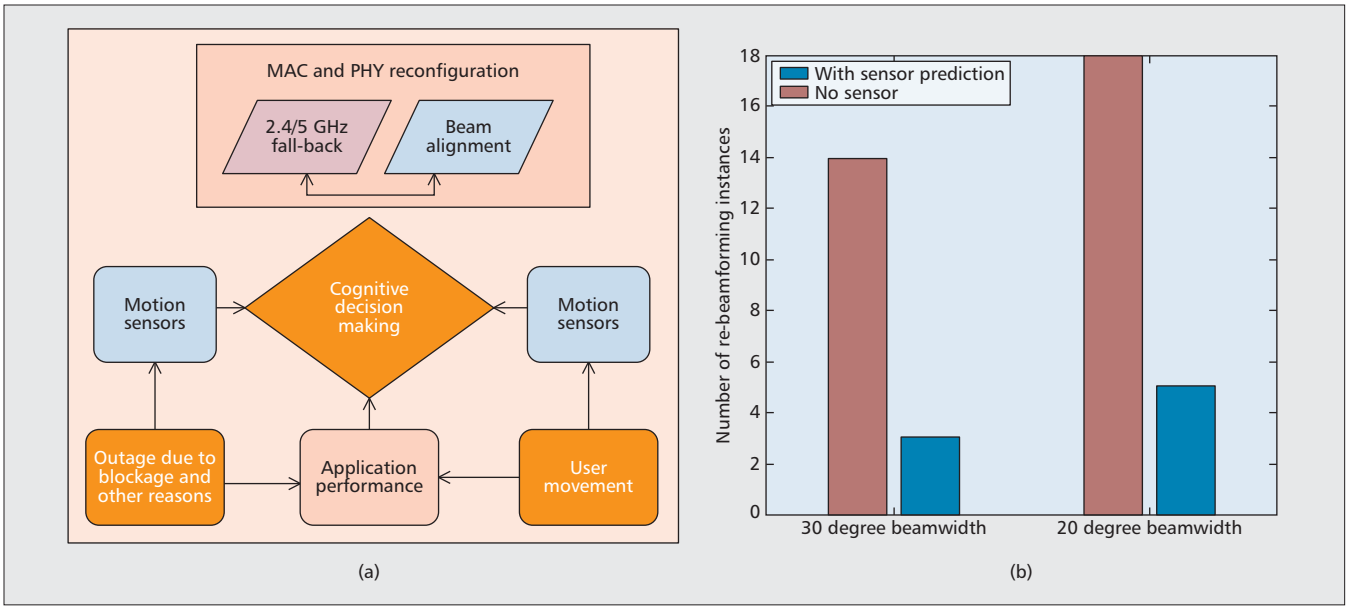


Figure 5. Sensor-assisted intelligent beamforming: a) system diagram of sensor assisted beamforming; b) re-beamforming instances.

[13] T. Nitsche *et al.*, "Steering with Eyes Closed: MM-Wave Beam Steering Without In-Band Measurement," *Proc. IEEE INFOCOM*, 2015.

[14] X. An, R. Venkatesha Prasad, and I. Niemegeers, "Impact of Antenna Pattern and Link Model on Directional Neighbor Discovery in 60 GHz Networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 5, May 2011, pp. 1435–47.

[15] A. W. Doff, K. Chandra, and R. V. Prasad, "Sensor Assisted Movement Identification and Prediction for Beamformed 60 GHz Links," *2015 IEEE 12th Consumer Communications and Networking Conference (CCNC) (CCNC 2015)*, Las Vegas, USA, Jan. 2015.

Biographies

Kishor Chandra (K.Chandra@tudelft.nl) is currently pursuing his Ph.D. in the Embedded Software group of Delft University of Technology, The Netherlands. He received his M.Tech. degree in signal processing from the Indian Institute of Technology, Guwahati, India in 2009, and his B.Eng. in electronics and communications engineering at K.E.C. Dwarahat (Kumaon University), Nainital, India in 2007. Prior to joining Ph.D., he was a research engineer working on IP multimedia subsystems with the Centre for Development of Telematics (CDOT), New Delhi, India. His research interests are in the area of 60 GHz communications, 5G networks, and millimeter wave radio-over-fiber networks.

R. Venkatesha Prasad (R.R.VenkateshaPrasad@tudelft.nl) received his bachelor's degree in electronics and communication engineering and the M.Tech degree in industrial electronics from the University of Mysore, India, in 1991 and 1994, respectively. He received a Ph.D. degree in 2003 from IISc. In 1996 he worked as a consultant and project associate for the ERNET Lab of ECE at IISc. While pursuing his Ph.D. degree, from 1999 to 2003 he also worked as a consultant for CEDT, IISc, Bangalore for VoIP application development as part of Nortel Networks sponsored project. In 2003 he headed a team of engineers at ESQUBE Communication Solutions for the development of various real-time networking applications. From 2005 to 2012 he was a senior researcher at the Wireless and Mobile Communications group, Delft University of Technology, working on the EU funded projects MAGNET/MAGNET Beyond and PNP-2008, and guiding graduate students. From 2012 onward he has been an assistant professor at the Embedded Software group at Delft University of Technology. He is an active member of TCCN and IEEE SCC41, and a reviewer of many IEEE transactions and Elsevier journals. He is on the Technical Program Committees of many conferences, including IEEE ICC, IEEE GLOBECOM, ACM MM, ACM SIGCHI. He was TPC Co-Chair of the CogNet Workshop in 2007, 2008, and 2009, and TPC Chair for E2Nets at IEEE ICC '10. He has also been running the PerNets workshop from 2006 with IEEE CCNC. He was the tutorial Co-Chair of CCNC 2009 and 2011, and Demo Chair of IEEE CCNC 2010. He is secretary of the IEEE ComSoc Standards Development Board and associate editor of *IEEE Transactions on Emerging Telecommunications Technologies*.

Bien Quang (quang.bien@gmail.com) received his Ph.D. degree from Delft University of Technology in 2014. He received his B.S. degree and the M.Sc. degree in electronics and telecommunications from Hanoi University of Technology, Vietnam, in 2001 and 2004, respectively. His research interests include billing, mobility, home networking, and performance analysis of various wireless technologies, e.g. IEEE 802.15.3 and 802.11.

I. G. M. M. Niemegeers (I.G.M.M.Niemegeers@tudelft.nl) received a degree in electrical engineering from the University of Ghent, Belgium, in 1970. In 1972 he received an M.Sc.E. degree in computer engineering and in 1978 a Ph.D. degree from Purdue University in West Lafayette, Indiana. From 1978 to 1981 he was a designer of packet switching networks at Bell Telephone Mfg. Cy, Antwerp, Belgium. From 1981 to 2002 he was a professor on the Computer Science and Electrical Engineering faculties of the University of Twente, Enschede, The Netherlands. From 1995 to 2001 he was scientific director of the Centre for Telematics and Information Technology (CTIT) of the University of Twente, a multi-disciplinary research institute on ICT and applications. From May 2002 until his retirement in 2012 he held the chair of Wireless and Mobile Communications at Delft University of Technology, where he headed the Telecommunications Department. He is currently a professor emeritus at Delft University of Technology. He was involved in many European research projects, including the EU projects MAGNET and MAGNET Beyond on personal networks, EUROPCOM on UWB emergency networks, and eSENSE and CRUISE on sensor networks. His present research interests are 5G wireless infrastructures, future home networks, ad hoc networks, personal networks, and cognitive networks. He has (co)authored close to 300 scientific publications and has coauthored a book on personal networks.

Cognitive Spectrum Access in Device-to-Device-Enabled Cellular Networks

Ahmed Hamdi Sakr, Hina Tabassum, Ekram Hossain, and Dong In Kim

ABSTRACT

Cognitive spectrum access (CSA) in in-band D2D-enabled cellular networks is a potential feature that can promote efficient resource utilization and interference management among coexisting cellular and D2D users. In this article, we first outline the challenges in resource allocation posed by the coexistence of cellular and D2D users. Next, we provide a qualitative overview of the existing resource allocation and interference management policies for in-band D2D-enabled cellular networks. We then demonstrate how cognition along with limited information exchange between D2D users and the core network can be used to mitigate interference and enhance spectral efficiency of both cellular and D2D users. In particular, we propose a CSA scheme that exploits channel sensing and interference-aware decision making at the D2D terminals. This CSA scheme at the D2D terminals is complemented by a D2D-aware channel access method at the cellular BSs. The performance gains of the proposed CSA scheme are characterized in terms of channel access probability for a typical D2D transmitter and spectral efficiencies for both cellular and D2D transmissions. Finally, potential research issues that require further investigation are highlighted.

INTRODUCTION

Device-to-device (D2D) communication enables nearby wireless devices to exploit their proximity and communicate directly with each other, bypassing their corresponding cellular base stations (BSs) [1, 2]. By enabling single-hop communication instead of dual-hop uplink (UL) and downlink (DL) communication, D2D communication improves the radio resource utilization at the BSs, and enhances the latency, spectral efficiency, and power consumption of D2D transmitters (TXs). Also, it offloads traffic from the cellular BSs and thus reduces congestion on radio resources used by cellular user equipments (CUEs). Potential commercial applications of D2D communication include localized social networking and data transfer, home automation, and commerce and advertising. Public safety is another application where local connectivity can be ensured in the absence of BSs or hazards at BSs.

Since D2D transmissions typically occur in proximity, D2D terminals are expected to discover their peers (or communicating partners), select spectrum, schedule transmissions, and perform power control while avoiding interference from/to cellular transmissions in a smart manner. For instance, a D2D user can perform licensed spectrum sensing (similar to a cognitive radio) to detect the idleness of a given channel. Moreover, the D2D user can sense the surrounding environment to obtain required channel state information (CSI), interference, mobility, and other information related to nearby wireless devices. Exploiting cognition in D2D communication thus empowers the D2D users to make autonomous decisions and adjust their transmit power, operating frequency, and spectrum access policy opportunistically. Consequently, cognitive spectrum access (CSA) in D2D networks paves the way to developing distributed resource management solutions with reduced signaling overhead and complexity [3].

This article first overviews the different scenarios of D2D communication in cellular networks from an implementation perspective. Then the challenges related to resource allocation and interference management in D2D-enabled cellular networks are discussed, followed by a qualitative overview of the existing centralized and distributed resource allocation approaches. A CSA scheme is then proposed to demonstrate the impact of cognition and prioritized spectrum access in D2D-enabled cellular networks. The performance gains of the CSA scheme are then analyzed quantitatively. Finally, several potential directions for future research are outlined.

Ahmed Hamdi Sakr, Hina Tabassum, and Ekram Hossain are with the University of Manitoba.

Dong In Kim is with Sungkyunkwan University, Korea.

This work was supported in part by a Strategic Project Grant (STPGP 430285) from the Natural Sciences and Engineering Research Council of Canada (NSERC) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (2014R1A5A1011478).

0163-6804/15/$25.00 © 2015 IEEE

D2D-Enabled LTE/LTE-A Networks

The Third Generation Partnership Project (3GPP) targets D2D communication in Long Term Evolution (LTE) Release 12 to provide new commercial or public safety proximity services (ProSe) [4]. In general, D2D communications can be enabled in a cellular network in three possible ways: *D2D-unaware* transmissions, *D2D-aware* transmissions, and *network-controlled D2D* transmissions.

In D2D-unaware transmissions, D2D users can exchange control and data packets with each other without any intervention from the eNB (i.e., BS). The eNB does not have any supervision over the radio resources used by the D2D pairs (e.g., power control and spectrum allocation), as shown in Fig. 1a. To be specific, coordination between D2D and cellular transmissions is not possible in this scenario. Note that the D2D users can use the PC5 interface, which is defined by the LTE standard for discovery and direct communication between D2D users. On the other hand, in D2D-aware transmissions, a D2D pair can perform limited information exchange with the core network through the eNB using the LTE-Uu (i.e., cellular link) interface. In this case, the core network can perform limited supervision of D2D transmissions for better coordination with the concurrent cellular transmissions. Nonetheless, this supervision should be limited in terms of information exchange and signaling overheads. The D2D users can possibly decide their mode of operation (i.e., D2D or cellular) and the radio resources for data transmission using the PC5 interface (as shown in Fig. 1b).

In network-controlled D2D transmissions, an eNB fully controls the radio resources management of all cellular and D2D users in a cell. The network architecture in this case is similar to the one shown in Fig. 1b but with full control exercised by the eNB. This scenario enables the network to make perfect coordination between cellular and D2D users, which may require a large amount of information exchange and signaling overheads. Moreover, a D2D pair cannot establish a communication link without initiating a request and the approval of the request from the eNB.

Figure 2 shows the control and data protocol stacks for the network-controlled D2D-enabled LTE/LTE-Advanced (LTE-A) cellular networks [5]. It can be seen in Fig. 2a that the LTE control plane is reused for the D2D control plane over the LTE-Uu interface, where there is no control signaling between D2D users. On the other hand, the D2D user plane in Fig. 2b reuses the LTE data protocol stack with the introduction of the PC5 interface. The same protocol stacks in Fig. 2 can be used for D2D-aware deployment with minimal use of the LTE-Uu interface. Finally, for D2D-unaware cellular networks, D2D users do not interact with the radio access network; hence, data and control signaling are performed over the PC5 interface between each D2D pair.

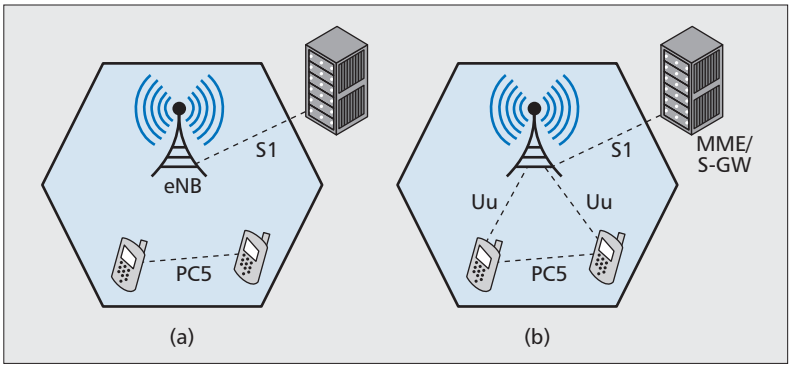


Figure 1. Architecture of a) D2D-unaware; b) D2D-aware LTE cellular networks.

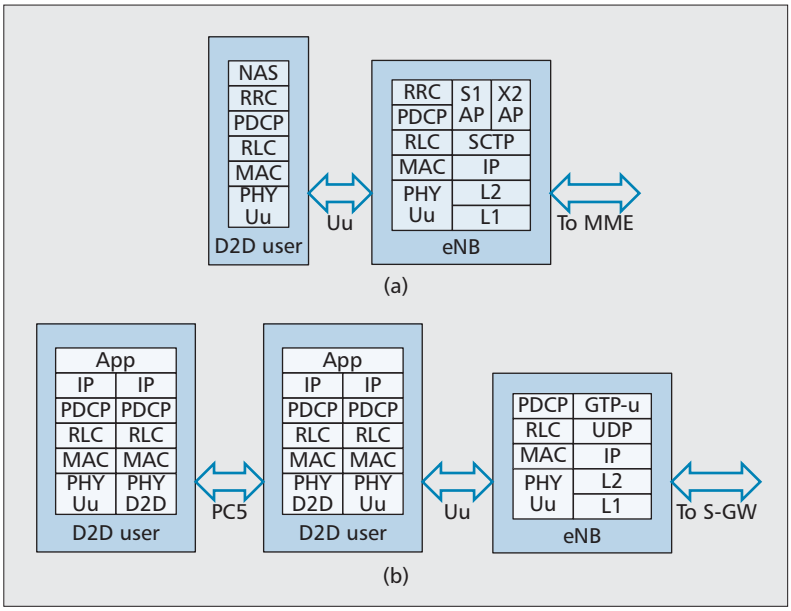


Figure 2. Protocol stacks for D2D-enabled LTE networks: a) control plane; b) user plane.

Fundamental Challenges of Resource Allocation in D2D-Enabled Cellular Networks

For in-band D2D communication (i.e., where the same radio resources are used for both cellular and D2D communication), the primary resource allocation challenges include interference management among coexisting CUEs and D2D users, development of low-complexity centralized or semi-distributed spectrum access and interference mitigation techniques with minimal signaling overheads, and the notion of priority for the resource management of D2D and CUEs. These challenges are elaborated below.

Resource Allocation

Network-controlled D2D architecture considers centralized resource allocation where the BS in a cell allocates resources to both CUEs and D2D users to improve its own utility. However, the centralized approaches may incur high computational and signaling complexities. For instance,

For high-intensity CUEs, prioritizing spectrum for D2D transmissions will not work. As such, simple D2D-aware scheduling techniques need be developed that allocate resources to D2D users in an opportunistic manner while providing fairness among CUEs and D2Ds.

the BS may need to be totally aware of the interference status at various D2D pairs or the channel information between a D2D pair for efficient spectrum allocation. In contrast, distributed resource allocation methods could be simpler. For example, local sensing can be used at the D2D terminals to sense the network environment and adaptively utilize the radio resources. Local sensing will enable the D2D terminals to measure the harmful interference (e.g., from CUEs and BSs, respectively, when uplink and downlink resources are used) and utilize this information to improve the spectral efficiency of D2D transmission. Another approach could be to exploit message passing [6]. This message passing approach requires exchange of local information among neighbors. However, in a densely deployed multi-tier network, the signaling overhead can become very high. As such, exploiting cognition at D2D terminals with limited information exchange with the BSs can be useful in such scenarios.

Prioritization of Cellular and D2D Users

In D2D-enabled cellular networks, typically, CUEs are offered a higher priority than D2D users. For example, a typical assumption in most of the existing literature is that D2D users can communicate only if they do not cause excessive interference to the CUEs. Since D2D communication aims at offloading traffic from cellular BSs in a distributed manner, priority for D2D users may need to be implemented in such a way that the performance of CUEs is not affected. For instance, to facilitate D2D communication, a few channels may not be assigned to the CUEs (i.e., allocated for D2D communication only) until the traffic load of CUEs becomes high enough to require those channels. This information can be transmitted by the BSs to the D2D pairs.

Cognitive Spectrum Access

Spectrum sharing between CUEs and D2D pairs allows higher spectrum reuse. However, it may lead to severe cross-tier interference at D2D links when they coexist with CUEs and other tiers such as the small cell tier in a multi-tier cellular network. On the other hand, static spectrum splitting among different tiers eliminates cross-tier interference but still could significantly degrade spectral efficiency depending on the number of D2D terminals and the proportion of available spectrum for them. Cognitive spectrum access methods with limited control of BSs can potentially adapt to the traffic load intensities of CUEs and D2D pairs.

D2D-Aware Scheduling in High Traffic Load

For high-intensity CUEs, prioritizing spectrum for D2D transmissions will not work. As such, simple D2D-aware scheduling techniques need be developed that allocate resources to D2D users in an opportunistic manner while providing fairness among CUEs and D2Ds. In this regard, D2D terminals can exploit cognition to sense user activity in a set of channels and then inform the BS about their most favorable channel. The BS then allocates the channels considering the overall intensity of D2D pairs and CUEs, and informs the D2D users about the spectrum allocation.

Management of Cross-Tier and Inter-D2D Interference

The introduction of D2D communication in cellular networks is challenging for both D2D and CUEs due to the cross-tier interference resulting from the concurrent cellular and D2D transmissions. This issue is more challenging in multi-tier networks in which low-power small cells are densely deployed over existing single-tier networks (as will be in the emerging fifth generation, 5G, cellular wireless networks) [7]. These small cells result in additional cross-tier interference on top of that from the macro BSs and CUEs. Efficient interference management techniques (e.g., power control, spectrum allocation, multiple antenna beamforming) will therefore be essential. Furthermore, the interference incurred at a D2D receiver from neighboring D2D transmitters (referred to as *inter-D2D interference*) also needs to be mitigated through proper user pairing and frequency assignment techniques.

The State-of-the-Art Resource Allocation Methods for In-Band D2D-Enabled Cellular Networks

A qualitative summary of the state-of-the-art approaches for centralized and distributed resource allocation in D2D-enabled cellular networks is provided in Table 1 when all D2D pairs and CUEs share the same radio spectrum.

Centralized Approaches

In [8], a near-optimal greedy resource allocation scheme is proposed for a single-cell scenario with multiple underlaying D2D pairs. The duration of simultaneous D2D transmissions is minimized while meeting cross-tier interference thresholds, signal-to-interference-plus-noise ratio (SINR) for D2D links, and maximum power constraints. The problem is formulated as mixed integer programming (MIP), and a column-generation-based method is proposed to obtain a low-complexity centralized solution where cellular links are given higher priority.

In [9], the authors consider a single-cell uplink network where multiple CUEs and one D2D pair can share the same radio spectrum. In this article, the cross-tier interference caused by cellular transmissions is managed to improve the overall network throughput where the interference from D2D transmissions is ignored, assuming that a power control method is in place. The interference management scheme does not allow CUEs to coexist with the D2D pair in the same spectrum resource if the resulting interference-to-signal-ratio (ISR) at the D2D receiver becomes higher than a predefined threshold.

A sum-rate maximization problem is formulated with both D2D and CUEs considering D2D-enabled uplink MIMO cellular networks

Approach	Article	Objective and assumptions	Overhead	Spectrum	User priority
Centralized	[8]	–RA for spectrum utilization maximization –Single cell and multiple D2D pairs –MIP solved by a column-generation-based method	Medium	DL/UL	Cellular
	[9]	–Interference management to enhance overall data rate –Single cell and one D2D pair, multiple CUEs –Scheme to control interference limited area is proposed	Low	UL	D2D
	[10]	–Network rate maximization with target rate constraints for both D2D and CUEs –Single cell with multiple CUEs and D2D pairs –Non-cooperative RA game is solved by self-optimization algorithm	High	UL	Both
Distributed	[6]	–Subcarrier and power allocation for two-hop rate maximization –Relay-assisted cellular network –Signaling overhead due to messages between relay and CUEs –Message passing approach	Low	UL	Both
	[11]	–Opportunistic spectrum access protocol with minimal SINR degradation of CUEs –Statistical estimates of the channel gains –Network information in the discovery packet	Medium	UL	Cellular
	[12]	–RA for D2D mode selection optimization and utility maximization –Utility of secondary users in D2D mode and total secondary users –Evolutionary game model	Medium	DL/UL	Cellular
	[14]	–Maximize D2D rate under CUEs' interference constraints –Overhead due to price broadcast at each channel –BS requires CSI of cellular links –D2D TX requires CSI of the D2D link and its link to the BS –Stackelberg game	Low	UL	Cellular
	[15]	–Network utility maximization with pricing –Near optimal solution –CSI of both D2D and CUEs required at BS –Reverse iterative combinatorial auction method	Medium	DL	N/A

Table 1. Existing literature on in-band D2D-enabled cellular networks.

[10]. Optimal resource allocations are obtained by using a pure random search. Moreover, a non-cooperative resource allocation game for joint channel allocation, power control, and precoding of D2D users is formulated. The feasibility and existence of the pure strategy Nash equilibrium are then established by developing a self-optimizing algorithm. Finally, a distributed resource allocation algorithm based on the best response dynamic is proposed.

Distributed Approaches

In [6], a distributed resource allocation scheme is proposed to maximize network sum rate while satisfying the data rate requirements for CUEs and D2D users considering a multi-user and multi-relay network. A message passing technique is used where each user sends and receives information messages to/from the relay node in an iterative manner with the goal of achieving an optimal allocation. The authors in [11] propose a distributed CSA policy in which D2D users can opportunistically share the UL spectrum resources with the CUEs. In order to limit the degradation of SINR in cellular links, each D2D user is assumed to perform power control to limit the level of cross-tier interference. Then the D2D user determines if the link should be established in a single- or multihop manner. A random access technique is used to mitigate inter-D2D interference by allowing only one D2D pair to access the spectrum at a time.

In [12], an efficient resource allocation method is proposed for a cognitive cellular network with D2D communication. While a secondary user may operate in either cellular mode or D2D mode, the theory of evolutionary game is used to model the behavior of such users. The proposed solution attempts to optimize the mode selection criterion with a set of utilities considering the data rate, transmit power, and cross-tier interference. The application of several game models in the distributed resource allocation of D2D-enabled cellular networks is discussed in [13]. For D2D communication, noncooperative and auction game models are suggested as the most suitable option to solve the resource allocation problems.

An iterative two-stage pricing-based algorithm is proposed in [14] in which the BSs send a pricing signal depending on the gap between the aggregate interference from D2D links and a

CSA in D2D-enabled cellular networks can be realized by enabling channel sensing and interference-aware decisions at the D2D terminals distributively. To enhance the performance of CSA, a prioritized channel access policy for D2D transmissions is also used at the BSs, which offers limited network control.

predefined interference tolerance level. Next, each D2D link independently maximizes its utility consisting of a reward equal to its expected rate and a penalty proportional to the interference caused by this link to the BS, as measured by the pricing signal. In [15], a reverse iterative combinatorial auction method is used to optimize the system sum utility. A non-monotonic descending price auction algorithm is proposed to maximize the utility function that accounts for the channel gain from D2D and the costs for the system. The scheme is cheat-proof and converges in a finite number of iteration rounds.

The centralized or near-optimal solutions are not scalable for dense networks with large numbers of BSs, CUEs, and D2D users. On the other hand, the distributed solutions that exploit cognition at D2D terminals for automated channel sensing and decision will be scalable for such networks.

Cognitive Spectrum Access for D2D-Aware Cellular Networks

CSA in D2D-enabled cellular networks can be realized by enabling channel sensing and interference-aware decisions at the D2D terminals distributively. To enhance the performance of CSA, a prioritized channel access policy for D2D transmissions is also used at the BSs, which offers limited network control. To evaluate the performance of the proposed solution, we characterize:

- Channel access probability (CAP) of a D2D TX
- Spectral efficiency for both cellular and D2D transmissions.

Cognition at D2D Devices

A cognitive D2D user is capable of sensing the received interference level on a given transmission channel. With this knowledge, a D2D user can make an intelligent decision about utilizing a given channel while avoiding interference from nearby cellular transmissions (in UL or DL). We exploit cognition at the D2D RXs. That is, the D2D TX sends a ready-to-send (RTS) request to its intended receiver over a PC5 interface. Then, on a given channel, if the maximum received interference from any neighboring TX is sensed to be lower than a predefined sensing threshold γ, the intended D2D RX sends a clear-to-send (CTS) signal over the PC5 interface to its corresponding TX.

Cognition at the D2D RXs provides a protection region around each D2D RX in which a D2D communication link cannot be established if there is at least one TX (i.e., CUE in uplink or BS in downlink) using the same channel inside this region. In general, the protection region around each D2D RX has a random shape due to the randomness in channel conditions and transmit power. As an example, for channel gain h and an interferer at a distance r with transmit power P, the decision is made by comparing the received interference power Phr^{-4} (assuming that the path loss exponent for radio propagation is 4) to the sensing threshold γ. If $r > (Ph/\gamma)^{1/4}$, this means that the interferer is outside the protection region, and the D2D TX can thus transmit; otherwise, the D2D TX remains silent. Note also that decreasing γ provides more protection to the D2D RXs by decreasing the aggregate interference; however, it reduces the channel access probability for the D2D TXs. To avoid inter-D2D interference, we assume that, after sensing a free channel, each D2D TX sets a random backoff timer upon expiration of which the D2D TX can use the channel if the channel is still available.

Spectrum Access Policy Adopted by BSs

We consider the two following spectrum access techniques at the cellular BS:

- D2D-unaware spectrum access (DUSA)
- D2D-aware spectrum access (DASA)

as described below. The first policy is the baseline scenario in which BSs assign different frequencies to their users without considering the D2D transmissions. On the other hand, the second policy is a conservative policy in which the BSs avoid causing interference to the D2D transmissions whenever possible.

DUSA policy: A BS can utilize any channel to serve any CUE based only on its scheduling policy, whereas each D2D pair selects a random channel to perform CSA.

DASA policy: A BS first assigns one of the channels (say c_d) for D2D transmissions and exchanges the ID of this channel with the D2D users in its coverage area over the LTE-Uu interface. Once the D2D users are informed of the D2D channel, they are responsible for initiating the communication session by exchanging control and data signals over the PC5 interface with no further supervision from the BS. This channel is not exclusive for D2D users and can be used for cellular transmissions based on the following policy: The BS schedules CUEs in any of its available channels except c_d as long as the number of CUEs N_u is less than the total number of available channels C. Thus, c_d is guaranteed to be the least congested channel. That is, the D2D transmissions can fully exploit c_d when $N_u < C$. When $N_u \geq C$, a CUE and a D2D user will compete for c_d. In this case, the CUE is granted the channel since it normally has higher priority than D2D users.

Using the DASA policy with CSA (referred to as the DASA-CSA policy), the interference at D2D RXs can be further minimized while improving the overall spectral efficiency of D2D transmissions. To compare the performance of the DASA and DUSA policies, we calculate the probability for a D2D pair to find a free channel to establish a communication link. The channel access probability in a given cell is directly related to N_u. Users are uniformly distributed, and N_u is modeled by a Poisson random variable with mean $\sqrt{3}/2\mathcal{U}d^2$, where $\mathcal{U}$ is the spatial intensity of CUEs and d is the inter-BS distance. This is derived by using the area of the hexagonal cell $A = \sqrt{3}/2\ d^2$.

To cope with the dynamics of spectrum access in a D2D-enabled cellular network with the DASA-CSA policy, the D2D user can exchange other information such as the spectrum sensing range (or equivalently, the spectrum sensing threshold) for the cognitive D2D transmitters, with the BS over the PC5 interface.

Performance Evaluation Results

We compare the four possible scenarios: DUSA-only without CSA, DASA-only without CSA, DUSA-CSA, and DASA-CSA, in terms of CAP and spectral efficiency. The results are shown in Fig. 3, which shows the effect of increasing the spatial intensity of CUEs on the CAP for a D2D TX. Note that the CAP for the DASA-CSA scenario is equivalent to the probability that the number of CUEs served by a BS is at most $C - 1$, that is,

$$\text{CAP of DASA-CSA} = \sum_{n=0}^{C-1} \mathbb{P}\{N_u = n\},$$

where $\mathbb{P}\{N_u = n\}$ is the probability mass function of N_u which is a function of the spatial intensities of BSs and CUEs and the user association policy.

On the other hand, for the DUSA-CSA scenario, the CAP is obtained using the fact that each BS assigns all channels to different users with the same probability, that is,

$$\sum_{n=0}^{C-1} (1 - \frac{n}{C}) \mathbb{P}\{N_u = n\}.$$

In this comparison, the network has a total of 15 channels where the inter-site distance is 500 m according to the 3GPP evaluation methodology.

Impact of cognition (CSA vs. no-CSA): It can be seen in Fig. 3 that exploiting cognition in D2D transmissions can greatly improve the CAP for a D2D TX for both DUSA and DASA scenarios. Note that cognition is more advantageous in dense networks due to significant interference. For example, with the DUSA policy, the improvement in CAP due to cognition is only 3 percent (i.e., from 0.85 to 0.87) when the density of CUEs is 10 CUEs/km^2 compared to 85 percent (i.e., from 0.29 to 0.53) when the density of CUEs is 50 users/km^2. This result suggests that CSA is not crucial when the number of CUEs is low since not all radio resources are used by the macro-tier and the D2D TXs have a good chance of finding free channels. On the other hand, for dense networks, CSA is crucial to avoid the nearby interferers and increase the efficiency of using the available resources.

Impact of D2D-awareness (DUSA vs. DASA): Figure 3 also quantifies the performance gain of DASA over the DUSA policy for both scenarios (i.e., with and without CSA). The figure shows that as the density of CUEs increases, CAP degrades for both DUSA and DASA. For example, increasing the CUEs in the DUSA-only scenario from 10 to 50 users/km^2 degrades CAP by a factor of 3 (i.e., from 0.85 to 0.29). Most importantly, it can be seen that with and without cognition, the DASA scenario offers better performance for all network parameters compared to the DUSA scenario. For instance, in the D2D-aware scenario, with a 5× increase in the intensity of CUEs from 10 to 50 users/km^2, the CAP becomes 0.87 compared to only 0.29 in the D2D-unaware scenario. In addition, we can see that the gains of DASA become more evident in dense networks (i.e., the CAP decreases significantly). Note that this improvement comes at the expense of making other channels more congested for the CUEs. However, in the worst case, when all BSs have more users than the available channels, both access policies offer similar performance.

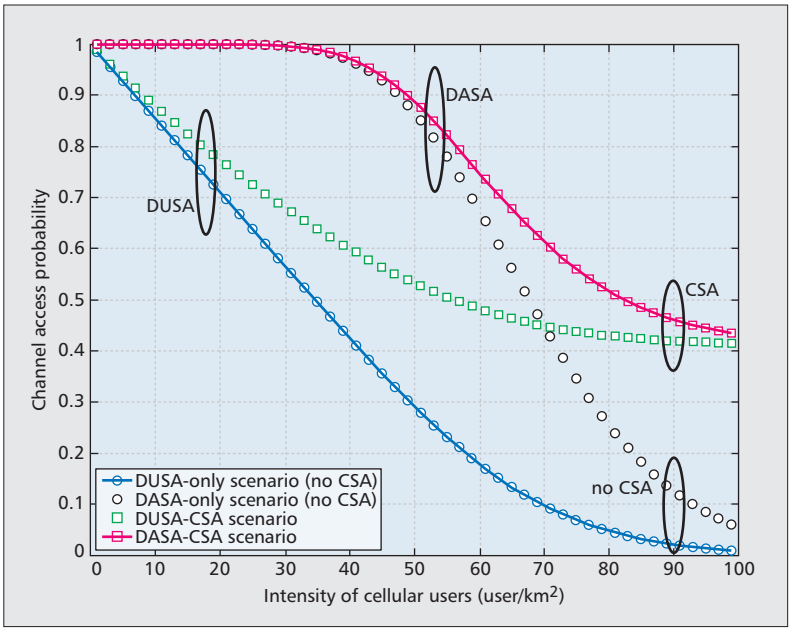


Figure 3. Channel access probability for a D2D TX vs. spatial intensity of cellular users for DUSA and DASA policies (with and without cognition).

Gain of the proposed framework (DASA-CSA vs. DUSA-only): It can be seen from Fig. 3 that there is a value for the spatial intensity of CUEs below which the gain of prioritized spectrum access is higher than that of CSA. After this point, using CSA is more beneficial. Therefore, combining both the schemes will offer better performance for high and low spatial intensities. Figure 3 shows that the scenario in which both cognition and prioritized techniques are combined (i.e., DASA-CSA) gives superior performance compared to the baseline scenario (DUSA-CSA) for all values of $\mathcal{U}$. For instance, in the scenario with 70 CUEs/km^2, DASA-CSA offers a 525 percent improvement in performance in terms of CAP for D2D TXs.

Furthermore, Fig. 4 illustrates the gain in spectral efficiency (SE) when combining cognition and prioritization in D2D-enabled networks. In this context, the spectral efficiency is defined as the number of successfully transmitted bits per unit time per Hertz where the data is considered successfully transmitted when the received signal-to-interference-plus-noise ratio (SINR) is higher than a predefined threshold τ. The following remarks can be made.

- The spectral efficiency for both cellular and D2D users degrades with increasing $\mathcal{U}$. This is intuitive due to the increase in both cross-tier and co-tier interference levels.
- The DASA-CSA policy can provide higher data rates for D2D transmissions compared to the DUSA-only scenario, especially for dense

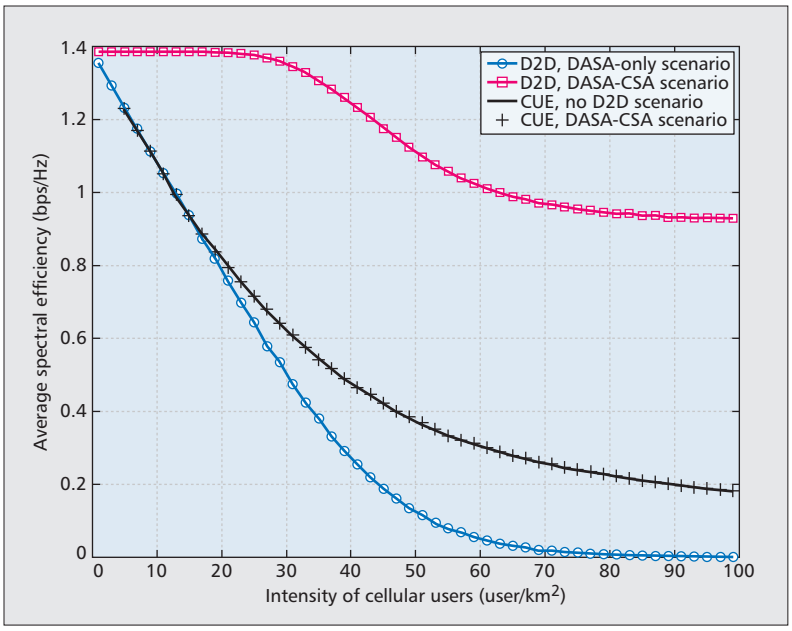


Figure 4. Spectral efficiency of transmission in D2D links vs. spatial intensity of cellular users for DUSA and DASA policies (with and without cognition).

cellular networks. For example, the DUSA-only scenario cannot support an SE of 1 b/s/Hz for D2D transmissions when the number of CUEs is higher than 13 users/km^2, while the DASA-CSA scenario provides the same SE with minimal coordination for CUE intensity up to 63 CUEs/km^2.

- The proposed DASA-CSA scheme does not impact the performance of CUEs where the gap between the SE of transmission by cellular users in this scenario, and the scenario with no D2D network is negligible.
- While the SE of CUEs is higher than that of the D2D users in the DUSA-only scenario due to the cross-tier interference, the opposite happens in the DASA-CSA scenario since the D2D users almost operate in an interference-free environment using their cognition capabilities.

Open Research Issues

The performance of cognitive spectrum access in D2D-enabled cellular networks can be further improved (e.g., in terms of both spectrum efficiency and energy efficiency) by using more advanced resource allocation methods as well as emerging communication techniques such as full-duplexing, and also RF energy harvesting techniques. In this context, several research directions are outlined below.

1) Traffic-load-aware spectrum selection: Due to highly asymmetric UL and DL data traffic in cellular networks, spectrum selection for D2D transmission needs to be adaptive to the traffic load. Traffic-load-aware channel selection has a direct impact on the interference, spectral efficiency, and receiver complexity.

2) Optimal spectrum allocation for prioritized D2D channels: Since the proposed CSA scheme along with BS awareness of D2D communication allocates part of the spectrum for D2D transmissions, an important question is how to optimally decide the proportion of prioritized resources for D2D transmissions. Given the intensity of D2D and CUEs, the proportion of prioritized spectrum needs to be decided in a dynamic fashion.

3) Admission control to manage inter-D2D interference: The performance gains of D2D transmissions highly depend on the intensity of D2D pairs on a given channel and their distances from each other. Efficient D2D admission control algorithms need to be developed that maximize the overall utility of both the CUEs and the D2D users. D2D admission control can be triggered in an opportunistic manner and the transmission of D2D users can be switched through the BS when communication using the direct D2D links is no longer beneficial.

4) Exploiting full-duplex transmissions: While D2D links are typically exploited for half-duplex data transmissions, they may be used for interference mitigation or channel selection if exploited in full-duplex mode. For instance, a cognitive D2D transmitter, when operating in full-duplex mode, can hear interference signals and provide information about the interference to its intended receiver along with the data packets. In a similar way, a CUE receiving transmission in the DL can also exploit the D2D communication to forward interference knowledge to a nearby cognitive D2D user. This knowledge can help the D2D user in either interference cancellation or channel selection.

5) Cognitive spectrum access in the presence of RF energy harvesting: With cognitive spectrum access, a D2D user can harvest energy while performing spectrum sensing or waiting for a free channel. Then the stored energy can be used later for data transmissions. Thus, resource allocation schemes need to be aware of the amount of energy available at the transmitter. Also, the duration of data transmission and energy harvesting will need to be optimized to maximize the system performance and reliability.

6) Combination of full-duplex and energy harvesting: If a D2D transmitter is equipped with dual antennas, one for transmission and the other for reception, the hybrid mode of information transmission and energy harvesting can be implemented using one antenna for each purpose. Unlike the information reception in traditional full-duplex mode, the self-interference from the transmission in full-duplex mode can be utilized for energy harvesting since decoding of the self-interference would not be required in this case. Therefore, adaptive (or cognitive) D2D can be used for this hybrid mode to further enhance the spectrum efficiency and energy efficiency through energy harvesting in full-duplex mode instead of the costly self-interference cancellation for information reception.

Conclusion

This article has focused on highlighting the key challenges in the resource allocation of in-band D2D-enabled cellular networks. A qualitative overview of the existing research advancements related to centralized and distributed resource allocation techniques has been provided. Since

centralized solutions generally incur high computation and signaling overheads, distributed or semi-distributed solutions that exploit cognition at the D2D terminals are promising. We have proposed a semi-distributed CSA solution in which cognition at the D2D terminals allows interference-aware decision making, and limited control at the BSs helps the D2D users select the spectrum band with the least interference. The performance of the CSA scheme has been analyzed quantitatively in terms of channel access probability and spectral efficiency of transmission in cellular and D2D links. The performance of CSA in a D2D-enabled cellular network can be optimized through a proper choice of network design parameters such as the spatial intensity of BSs, the mode selection criteria, and the spectrum sensing threshold.

References

[1] K. Doppler *et al.*, "Device-to-Device Communication as an Underlay to LTE-Advanced Networks," *IEEE Commun. Mag.*, vol. 47, no. 12, Dec. 2009, pp. 42–49.

[2] D. Feng *et al.*, "Device-to-Device Communications in Cellular Networks," *IEEE Commun. Mag.*, vol. 52, no. 4, Apr. 2014, pp. 49–55.

[3] E. Hossain, D. Niyato, and Z. Han, *Dynamic Spectrum Access and Management in Cognitive Radio Networks*, Cambridge Univ. Press, 2009.

[4] 3GPP TR 23.703 v12.0.0, "Technical Specification Group Services and System Aspects: Study on Architecture Enhancements to Support Proximity-Based Services (ProSe)," Rel 12," Feb. 2014.

[5] L. Song *et al.*, *Wireless Device-to-Device Communications and Networks*, Cambridge Univ. Press, 2015.

[6] M. Hasan and E. Hossain, "Distributed Resource Allocation for Relay-Aided Device-to-Device Communication: A Message Passing Approach," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, Nov. 2014, pp. 1536–76.

[7] E. Hossain, L. B. Le, and D. Niyato, *Radio Resource Management in Multi-Tier Cellular Wireless Networks*, Wiley, 2013.

[8] P. Phunchongharn, E. Hossain, and D. I. Kim, "Resource Allocation for Device-to-Device Communications Underlaying LTE-Advanced Networks," *IEEE Wireless Commun.,* vol. 20, no. 4, Aug. 2013, pp. 91–100.

[9] H. Min *et al.*, "Capacity Enhancement Using an Interference Limited Area for Device-to-Device Uplink Underlaying Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, Dec. 2011, pp. 3995–4000.

[10] W. Zhong *et al.*, "Joint Resource Allocation for Device-to-Device Communications Underlaying Uplink MIMO Cellular Networks," *IEEE JSAC*, vol. 33, no. 1, Nov. 2014, pp. 41–54.

[11] B. Kaufman, J. Lilleberg, and B. Aazhang, "Spectrum Sharing Scheme between Cellular Users and Ad-Hoc Device-to-Device Users," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, Mar. 2013, pp. 1038–49.

[12] P. Cheng *et al.*, "Resource Allocation for Cognitive Networks with D2D Communication: An Evolutionary Approach," *Proc. IEEE WCNC*, 2012, pp. 2671–76.

[13] L. Song *et al.*, "Game-Theoretic Resource Allocation Methods for Device-to-Device Communication," *IEEE Wireless Commun.*, vol. 21, no. 3, June 2014, pp. 136–44.

[14] Q. Ye *et al.*, "Distributed Resource Allocation in Device-to-Device Enhanced Cellular Networks," *IEEE Trans. Commun.* vol. 63, no. 2, Dec. 2014, pp. 441–54.

[15] C. Xu *et al.*, "Efficiency Resource Allocation for Device-to-Device Underlay Communication Systems: A Reverse Iterative Combinatorial Auction based Approach," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 348–58.

Biographies

AHMED H. SAKR [S'12] (Ahmed.Sakr@umanitoba.ca) is a Ph.D. candidate in the Department of Electrical and Computer Engineering, University of Manitoba, Canada (http://home.cc.umanitoba.ca/~sakra.html). He received his B.Sc. (2002–2007) and M.Sc. (2010–2012) degrees, both in electronics and communications engineering from Tanta University, Egypt, and the Egypt-Japan University of Science and Technology (E-JUST), Alexandria, Egypt, respectively. For his academic excellence, he has received several academic awards including the Manitoba Graduate Scholarship (MGS) in 2014, the Edward R. Toporeck Graduate Fellowship in Engineering in 2014, the Graduate Enhancement of Tri-Council Stipends (GETS) in 2013, and Egyptian Ministry of Higher Education Excellence Scholarship in 2010–2012. He has been a member in the technical program committee and a reviewer in several IEEE journals and conferences. His current research interests include statistical modeling of wireless networks, resource allocation in multi-tier cellular networks, and green communications.

HINA TABASSUM (Hina.Tabassum@umanitoba.ca) received her B.E. degree in electronic engineering from the NED University of Engineering and Technology (NEDUET), Karachi, Pakistan, in 2004. During her undergraduate studies she received two gold medals from NEDUET and Siemens for securing the first position among all engineering universities of Karachi. She then worked as a lecturer at NEDUET for two years. In September 2005, she joined the Pakistan Space and Upper Atmosphere Research Commission (SUPARCO), Karachi, and received there the best performance award in 2009. She completed her Master's and Ph.D. degrees in communications engineering from NEDUET in 2009 and King Abdullah University of Science and Technology (KAUST), Makkah Province, Saudi Arabia, in May 2013, respectively. Currently, she is working as a post-doctoral fellow at the University of Manitoba (UoM), Canada. Her research interests include wireless communications with focus on interference modeling, spectrum allocation, and power control in heterogeneous networks.

EKRAM HOSSAIN [F'15] (Ekram.Hossain@umanitoba.ca) is a professor (since March 2010) in the Department of Electrical and Computer Engineering at UoM. He received his Ph.D. in electrical engineering from the University of Victoria, Canada, in 2001. His current research interests include design, analysis, and optimization of wireless/mobile communications networks, cognitive radio systems, and network economics. He has authored/edited several books in these areas (http://home.cc.umanitoba.ca/~hossaina). He serves as Editor-in-Chief for *IEEE Communications Surveys and Tutorials* and an Editor for *IEEE Wireless Communications*. Also, he is a member of the IEEE Press Editorial Board. Previously, he served as an Area Editor for *IEEE Transactions on Wireless Communications* on Resource Management and Multiple Access from 2009 to 2011, an Editor for *IEEE Transactions on Mobile Computing* from 2007 to 2012, and an Editor for the *IEEE Journal on Selected Areas in Communications — Cognitive Radio Series* from 2011 to 2014. He has won several research awards including the University of Manitoba Merit Award in 2010 and 2014 (for Research and Scholarly Activities), the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award, and the IEEE Wireless Communications and Networking Conference 2012 Best Paper Award. He is a Distinguished Lecturer of the IEEE Communications Society (2012–2015). He is a registered Professional Engineer in the province of Manitoba, Canada.

DONG IN KIM [S'89, M'91, SM'02] (dikim@skku.ac.kr) received his Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1990. He was a tenured professor with the School of Engineering Science, Simon Fraser University, Burnaby, British Columbia, Canada. Since 2007, he has been with Sungkyunkwan University, Suwon, Korea, where he is currently a professor with the College of Information and Communication Engineering. He has served as an Editor and a Founding Area Editor of Cross-Layer Design and Optimization for *IEEE Transactions on Wireless Communications* from 2002 to 2011. From 2008 to 2011, he served as Co-Editor-in-Chief for the *Journal of Communications and Networks*. He is currently the Founding Editor-in-Chief for *IEEE Wireless Communications Letters* and has been serving as an Editor of Spread Spectrum Transmission and Access for *IEEE Transactions on Communications* since 2001. He was the recipient of the Engineering Research Center (ERC) for Wireless Energy Harvesting Communications Award.

We have proposed a semi-distributed CSA solution in which cognition at the D2D terminals allows interference-aware decision making and limited control at the BSs helps the D2D users in selecting the spectrum band with the least interference.

NETWORK AND SERVICE MANAGEMENT

George Pavlou

Jürgen Schönwälder

This is the 19th issue of the series on Network and Service Management, which is published twice a year. Until 2012 it was published in July and December, but since 2013 it has been published in January and July. The series provides articles on the latest developments in this well established discipline, highlighting recent research achievements and providing insight into both theoretical and practical issues related to the evolution of the discipline from different perspectives. The series provides a forum for the publication of both academic and industrial research, addressing the state of the art, theory, and practice in network and service management.

An important event for both the network/service management community but also for the networking research community as a whole has been the appointment of Prof. Aiko Pras of the University of Twente, Netherlands, as Chair of IFIP TC6 Communication Systems, taking over from Prof. Guy Leduc of the University of Liege, Belgium. Prof. Pras has also been Chair of IFIP 6.6 Management of Networks and Distributed Systems since 2008. Prof. Pras was also a Co-Editor of this series since its inception, having stepped down approximately a year ago.

The key annual event in this area, the IEEE/IFIP Integrated Management Symposium (IM 2015), was held May 11–15 in Ottawa, Canada; http://ieee-im.org/2015/. The second key annual event in this area will be the 11th International Conference on Network and Service Management (CNSM 2015), taking place November 09–13 in Barcelona, Spain; http://cnsm-conf.org/2015/. CNSM is a single-track flagship event sponsored and supported by IFIP, IEEE, and ACM. It complements IM and NOMS, which alternate every other year. The next edition of NOMS, NOMS 2016, will take place in Istanbul, Turkey, in April 2016. In addition, the 1st IEEE Conference on Network Softwarization (NetSoft 2015) was held April 13–17 in London, United Kingdom; http://sites.ieee.org/netsoft/. Finally, the European Conference on Autonomous Infrastructure Security and Management (AIMS 2015), which is supported by the European FLAMINGO project (see below), was held June 22–25 in Ghent, Belgium; http://www.aims-conference.org/2015/.

The European Network of Excellence FLAMINGO on the Management of the Future Internet (http://fp7-flamingo.eu/), which started at the end of 2012, has continued its activities successfully into a third year. The FLAMINGO project effectively continues the work that the EMANICS project (http://www.emanics.org/) started.

We again experienced excellent interest in the 19th issue, having received 16 submissions in total. For each paper we got at least three independent reviews. We finally selected four articles, resulting in an acceptance rate of 25 percent. It should be mentioned that the acceptance rate for all the previous issues has ranged between 14 and 25 percent, making this Series a highly competitive place to publish. We intend to maintain our rigorous review process in future issues, and thus the high quality of the published articles.

The first article, "Toward a Holistic Federated Future Internet Experimentation Environment: The Experience of NOVI Research and Experimentation" by Maglaris *et al.*, presents the key results of the European NOVI project on the issues behind a federated virtualization infrastructure for conducting realistic experiments related to future Internet research.

The second article, "Quality of Experience in Mobile Cellular Networks" by Liotou *et al.*, aims to provide insights on the issue of network-level QoE management in cellular networks, proposing a framework for end-to-end QoE provisioning and describing its design, constituents, and relevant interactions, as well as the key implementation challenges.

The third article, "An Open Framework for Programmable Self-Managed Radio Access Networks" by Poulios *et al.*, presents an approach for applying SDN principles to RANs, describing the architecture and implementation of a prototype for a programmable self-managed LTE-Advanced heterogeneous network.

Finally, the fourth article, "On Demand Scheduling: Achieving QoS Differentiation for D2D Communications" by Sheng *et al.*, first reviews recent research on relevant scheduling mechanisms and then proposes a new scheduling mechanism, DO-Fast, which can provide QoS differentiation capabilities, evaluating its performance.

We hope that readers of this issue again find the articles informative, and we will endeavor to continue with similar issues in the future. We would finally like to thank all the authors who submitted articles to this Series, and the reviewers for their valuable feedback and comments on the articles.

BIOGRAPHIES

GEORGE PAVLOU (g.pavlou@ucl.ac.uk) is a professor of communication networks in the Department of Electronic and Electrical Engineering, University College London, United Kingdom, where he coordinates research activities in networking and network management. He received a Diploma in engineering from the National Technical University of Athens, Greece, and M.Sc. and Ph.D. degrees in computer science from University College London. His research interests focus on networking and network management, including aspects such as traffic engineering, quality of service management, policy-based systems, autonomic networking, information-centric networking, and software-defined networks. He has been instrumental in a number of European and U.K. research projects that produced significant results with real-world uptake and has contributed to standardization activities in ISO, ITU-T, and IETF. He has been the Technical Program Chair of several conferences and in 2011 received the Daniel Stokesbury award for "distinguished technical contribution to the growth of the network management field."

JÜRGEN SCHÖNWÄLDER (j.schoenwaelder@jacobs-universiy.de) is a professor of computer science at Jacobs University Bremen, Germany, where he leads the Computer Networks and Distributed Systems research group. He received a doctoral degree from Technische Universität Braunschweig, Germany. His research interests include network management, distributed systems, network measurements, embedded networked systems, and network security. He is an active member of the Internet Engineering Task Force (IETF) where he has edited more than 30 network management related specifications and standards. He has co-chaired the ISMS working group of the IETF and currently serves as Co-Chair of the NETMOD working group. Previously, he chaired the Network Management Research Group of the Internet Research Task Force (IRTF). He has been involved in several European research projects, and has served in various roles for IEEE and IFIP sponsored conferences. He currently serves on the Editorial Boards of the *Springer Journal of Network and Systems Management* and the *Wiley International Journal of Network Management*. Previously, he served on the Editorial Board of *IEEE Transactions on Network and Service Management*.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
SEMANTICS FOR ANYTHING AS A SERVICE

BACKGROUND

Services (including anything-as-a-service) are the buzz in the industry. Networks are morphing to utilize new technologies like network functions virtualization and software-defined networking that are changing the way services are ordered, configured, and monitored. To support the evolving infrastructure, new network and service management platforms need to support standard mechanisms for communication within and across administrative domains. In order to support on-demand, dynamic configuration and monitoring, both common application programming interfaces (APIs) and a common language that has agreed semantics are required. Standards bodies are using information and data modeling to describe the abstract representations and detailed structured data needed by the orchestrators and controllers in the ecosystem.

This Feature Topic addresses the standards usage in the industry and advancements in the area of information and data modeling that support the semantics needed for end-to-end service management. Comparing and contrasting the top-down vs. bottom-up approaches to API development is also invited. Solicited topics include (but are not limited to):

- Information modeling
- Data modeling
- Transforming information models to data models
- Service development life cycle aspects
- End-to-end service management frameworks
- Model-driven development
- Modeling tools
- Landscape of YANG models
- Survey of modeling work from industry groups
- Advances needed in network management protocols
- Interaction of open source and traditional industry fora and standards development organizations

SUBMISSIONS

Articles should be tutorial in nature and written in a style comprehensible to readers outside the specialty of the article. Authors must follow *IEEE Communications Magazine*'s guidelines for preparation of the manuscript. Complete guidelines for prospective authors can be found at http://www.comsoc.org/commag/paper-submission-guidelines. It is very important to note that *IEEE Communications Magazine* strongly limits mathematical content, and the number of figures and tables. Paper length should not exceed 4500 words. All articles to be considered for publication must be submitted through the IEEE Manuscript Central site (http://mc.manuscriptcentral.com/commag-ieee) by the deadline. Submit articles to the "March 2016/Semantics for Anything as a Service" category.

SCHEDULE FOR SUBMISSIONS

- Manuscript Submission Due: September 15, 2015
- Decision Notification: November 15, 2015
- Final Manuscript Due: January 1, 2016
- Publication Date: March 2016

GUEST EDITORS

Scott Mansfield
Ericsson Inc.
scott.mansfield@ericsson.com

Hing-Kam Lam
Alcatel-Lucent
kam.lam@alcatel-lucent.com

Nigel Davis
Ciena
ndavis@ciena.com

Yuji Tochio
Fujistu
tochio@jp.fujitsu.com

Toward a Holistic Federated Future Internet Experimentation Environment: The Experience of NOVI Research and Experimentation

V. Maglaris, C. Papagianni, G. Androulidakis, M. Grammatikou, P. Grosso, J. van der Ham, C. de Laat, B. Pietrzak, B. Belter, J. Steger, S. Laki, M. Campanella, and S. Sallent

ABSTRACT

This article presents the design and pilot implementation of a suite of intelligent methods, algorithms, and tools for federating heterogeneous experimental platforms (domains) toward a holistic Future Internet experimentation ecosystem. The proposed framework developed within the NOVI research and experimentation European collaborative effort, aims at providing a modular data, control, and management plane architecture that includes: an information model capturing the abstractions of virtualized resources residing in different yet interworking experimental platforms; resource mapping algorithms tackling the inter-domain virtual network embedding problem; mechanisms providing interoperability of monitoring tools; policy-based management services for role-based intra and inter-domain management policies; and data-plane stitching mechanisms to enable the composition of user-specific slices (baskets of virtual resources drawn from the federated substrate). The NOVI framework was deployed and validated in a combined testbed consisting of two dissimilar platforms: a private PlanetLab domain with resources interconnected over the public Internet; and FEDERICA, an infrastructure of virtual resources interconnected via dedicated networking facilities of European National Research and Education Networks and GÉANT. This pre-normative work is expected to contribute to bridging Future Internet experimental federations with interconnected cloud architectures and interworked public/private data-centers, adding value via its intelligent services, information models, and composite algorithms.

INTRODUCTION

Over the last decade virtualization has been a promising technology for overcoming the ossification of the Internet, by enabling experimentation on new paradigms and architectures for the Future Internet (FI). Essentially, it allows for a large set of interesting experiments to run simultaneously over FI experimental facilities, pushing new ideas further into real implementations and deployment. The NOVI (Networking innovations Over Virtualized Infrastructures)[1] vision stems from the realization that combining virtualized computing and network infrastructures into a federated framework will drive further the adoption of FI experimental facilities for large-scale and diverse FI experimentation. According to the NOVI concept, federated FI experimental infrastructures should cater to dynamic provisioning, control, and monitoring of user-defined baskets (slices) of virtual resources drawn from loosely coupled, dissimilar, administrative domains.

The NOVI framework provides a modular data, control, and management plane federation architecture, validated within an integrated experimental prototype, mounted on interconnected FI experimental facilities. This involves the implementation of early prototypes of methods, algorithms, and information models that enable users to seamlessly use virtualized resources within the federation of heterogeneous research testbeds. This federated environment needs to be managed and configured within a stitched substrate, in order to allow for efficient and dynamic instantiation of resources across administrative domains, as demonstrated for Cloud computing resources by the Reservoir[2] FP7 project. In this article we report NOVI's contributions on open issues related to federated virtual infrastructures at the data, control, and management planes. These challenges are summarized in the following.

Information Model: In order to facilitate deployment of slices over federated FI experimental facilities, the use of an *information model* (IM) is required, that will provide a common vocabulary of concepts and capabilities that exist

V. Maglaris, C. Papagianni, G. Androulidakis, and M. Grammatikou are with National Technical University of Athens.

P. Grosso, J. van der Ham, and C. de Laat are with Universiteit van Amsterdam.

B. Pietrzak and B. Belter are with Poznan Supercomputing & Networking Center.

J. Steger and S. Laki are with Eötvös Loránd Tudományegyetem.

M. Campanella is with Consortium GARR.

S. Sallent is with Universitat Politècnica de Catalunya.

[1] FP7 Project NOVI, http://www.fp7-novi.eu

[2] FP7 Project Reservoir, http://www.reservoir-fp7.eu

0163-6804/15/$25.00 © 2015 IEEE

within such a federation. An IM as such should support [1]:

- Virtualization concepts to cater to virtualized resources (e.g. virtual machines, logical routers).
- Vendor independence as virtualized infrastructures employ hardware and software from different manufacturers.
- Monitoring and measurement concepts, so that monitored metrics (e.g. CPU utilization, one-way delay, available bandwidth) and corresponding units are uniformly specified across the federation.
- Support of management policies for the federated environment.

The definition of the *NOVI ontology-based IM* has been guided by these requirements in order to enable policy-based provisioning, monitoring, and lifecycle management of virtualized resources from the federated environment.

•Monitoring of the federated experimental facilities is a primary requirement, both for the experimenters that need to monitor the state of their slices/virtual resources as well as for control and management plane services that require monitoring information for decision-making. However, due to the multi-domain nature of the environment it is not a trivial task to provide federated monitoring functionalities. The heterogeneity of the federated networks (including network/computing elements and monitoring tools) poses a major challenge. The NOVI Monitoring Framework enables interoperability of monitoring tools operating within member platforms of the federation of heterogeneous testbeds catering to monitoring at the physical as well as the virtual level.

•Resource discovery and mapping/allocation are fundamental steps in the process of creating virtual networks (VNs). The problem of mapping VNs to specific nodes and links in a (multi-domain) substrate network is commonly referred to as the virtual network embedding (VNE) problem or the topology embedding problem. The general embedding problem is computationally intractable. Resource discovery has attracted less research attention than resource allocation [2], despite its large impact on the actual allocation process. The NOVI federation is empowered with a semantic-based, distributed resource discovery and mapping framework; semantic web techniques were exploited to facilitate interoperability in the federation of heterogeneous testbeds, while a hierarchical framework has been prototyped for distributed allocation of physical to virtual resources across the federated substrate.

•Network stitching of virtual resources belonging in different domains should provide at minimum Layer 2/Layer 3 connectivity. The problem becomes more challenging with the existence of heterogeneous network resources that need to be managed and configured within the federated environment. Appropriate stitching mechanisms must be utilized allowing for transparent data-plane connectivity across dissimilar platforms involved in multi-domain slices. Toward that end, the NOVI stitching scheme, based on Ethernet over GRE, introduces a programmable switch called NSwitch based on the Open vSwitch[3] implementation. Note that a similar approach was recently demonstrated by Hayashi *et al*.[4]

•Policy-based management needs to be revisited, in order to address additional requirements posed by the federated environment, such as inter-domain policies that define inter-platform duties. NOVI facilitates policy-based management of federated FI experimental infrastructures, employing a flexible engineering approach to establish relations between the testbed providers. The NOVI policy service enforces domain-independent management policies for both intra-domain and inter-domain management purposes, where each member-platform of a NOVI federation is considered as a separate domain.

NOVI builds on top of the Slice-Based Federation Architecture (SFA[5]). The SFA federation approach grew primarily out of the PlanetLab experience, and has been widely adopted in the NSF GENI program, as the basis for its GENI API [3]. SFA was also adopted in major European testbed federation initiatives, supported by the EU FIRE Unit.[6] SFA defines the notion of a slice, which is a container abstraction for all the resources in a given experiment [4]. Researchers are associated with slices and use SFA tools to discover and include resources in their slice. NOVI complements SFA by providing the necessary abstraction of resources (e.g. by introducing an ontology based IM) that enables advanced context aware services (e.g. monitoring of federated slices, distributed semantic-based resource discovery, and intelligent resource mapping service, management policies for both intra-domain and inter-domain management purposes, etc.), and a uniform data plane stitching mechanism. These value-added services, validated within the NOVI prototype implementation, are described in the following sections.

The NOVI policy service enforces domain-independent management policies for both intra-domain and inter-domain management purposes, where each member-platform of a NOVI federation is considered as a separate domain.

The NOVI Federation Concept and Prototype Implementation

The NOVI data and control and management architecture consists of three different layers, as depicted in Fig. 1.

At the bottom layer, heterogeneous FI experimental platforms provide the means to instantiate virtual resources per user slice. In the case of the NOVI prototype implementation, two particular platforms where used:

- A private PlanetLab [5] domain with resources (slivers) aggregated within three geographically distributed sites (at NTUA in Athens, Greece; PSNC in Poznan, Poland; and ELTE in Budapest, Hungary), interconnected via the public Internet;
- FEDERICA [6], an infrastructure of virtual resources (hosted at GARR in Milan, Italy; PSNC in Poznan, Poland; DFN in Frankfurt, Germany; NTUA in Athens, Greece; and i2CAT in Barcelona, Spain), interconnected via Layer 2 circuits of European National Research & Education Networks (NRENs) and GÉANT.[7]

Two instantiated slices are illustrated, denoted by blue and red colors in the figure. Using the *NSwitch*, Layer 2 connectivity is dynamically

[3] Open vSwitch, http://www.openvswitch.org

[4] VNode Project, "Federation Architecture and Common API / Common Slice Definition v2.0", March 2014, https://nvlab.nakao-lab.org/Common_API_V2.0.pdf

[5] Slice Federation Architecture, v2.0, http://groups.geni.net/geni/attachment/wiki/SliceFedArch

[6] Future Internet Research and Experimentation (FIRE), http://ict-fire.eu/

[7] GÉANT — http://www.geant.net/

The NOVI IM uses a semantic web approach and it is complemented by data models that use the Web Ontology Language. This choice has been driven by the desire to support reasoning and context awareness, which in turn allow NOVI to create efficient and complex services with resources available within the federation.

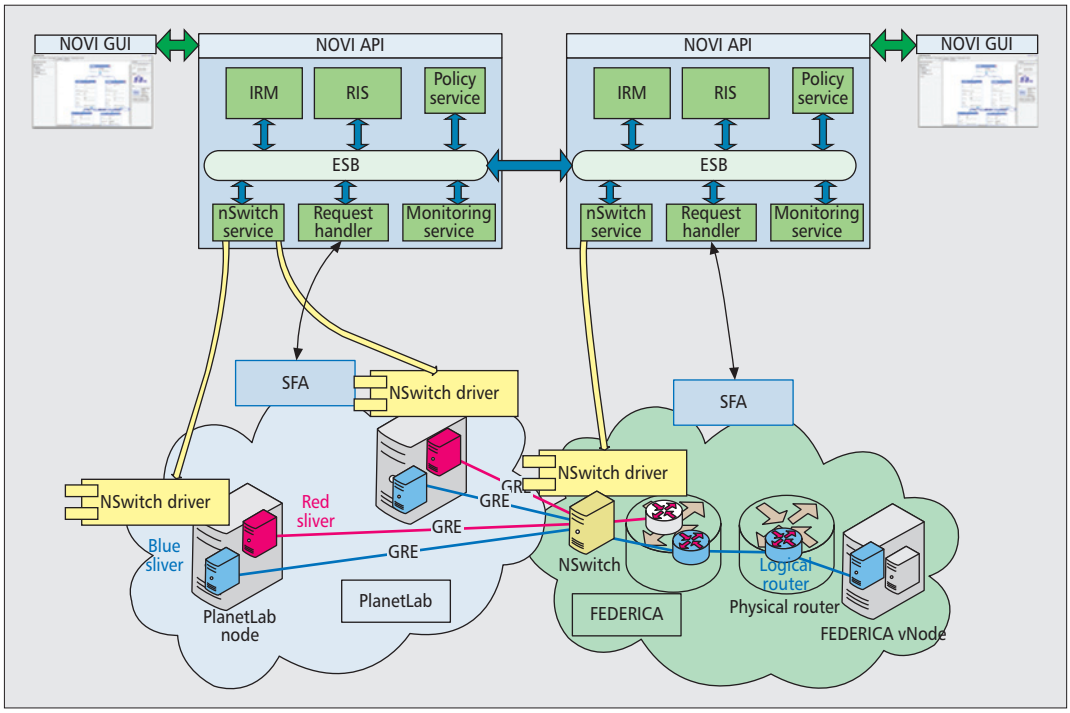


Figure 1. NOVI data, control, and management architecture.

established on demand, between virtual resources from different platforms that belong to the same slice. This is accomplished by means of Ethernet over GRE tunneling between PlanetLab nodes (via the NSwitch driver) and the NSwitch instance that is part of the FEDERICA testbed, allowing for native Ethernet transport within the FEDERICA domain.

At the middle layer components are used to provide basic control and management federation capabilities across platforms. In the figure, we depict implementation choices provided by the SFA, for example, cross-domain authentication via synchronized registries and user-specified slice operations; creation, instantiation, deletion; and so on.

The top layer, referred to as the NOVI service layer (SL), implements NOVI control and management services that offer advanced capabilities to the federation users, consistent with challenges summarized above. These are:

- The *resource information service* (RIS) responsible for context-aware resource discovery across the federation.
- The *intelligent resource mapping* service (IRM), mapping user requests for virtual resources to the federated physical substrate topologies.
- The *policy service* used to provide the functionality of a policy-based management system.
- The *NOVI monitoring service*, implementing the NOVI monitoring framework, that allows NOVI SL services as well as experimenters to retrieve monitoring information on physical and/or virtual resources across the federation.
- The *NSwitch* manager service communicating with NSwitch drivers deployed on platform resources, pushing configuration options based on the slice request.

The above intelligent services communicate with the underlying infrastructures via a *request handler* (RH) service that provides the NOVI SL with an abstraction of platform characteristics. In addition, the RH provides a generic implementation of an SFA request handler service, which is used to invoke the appropriate operations to the corresponding SFA APIs. Via the NOVI API, on top of the NOVI SL, experimenters can authenticate themselves, submit slice requests, view resources available at the federation, and monitor virtual resources. A small screenshot of the NOVI GUI is also shown in the figure, which allows users to formulate requests and submit them through the NOVI API. The NOVI SL integrated prototype implementation was based on the *Open Service Gateway initiative* (OSGi) framework[8] via an enterprise service bus (ESB). Services are implemented as OSGi bundles, loadable collections of classes and configuration files, thus providing modularity to the NOVI SL.

Each platform in the federation deploys a separate NSwitch, SFA, and NOVI SL. In what follows, we discuss functional and design specifications of architectural components within the above layers.

NOVI Architectural Components

Information Model for Federating Virtualized Infrastructures

The NOVI IM and the corresponding data models have a two-fold objective: to support the modeling abstractions to cater to a federation of the FEDERICA and PlanetLab platforms in the NOVI's testbed; and to include the necessary

[8] OSGi Service Platform Release 4, Version 4.2 Core Specification http://www.osgi.otg/Download/File?url=/download/r4v42/r4-core.pdf

concepts to model FI infrastructures that could participate in a NOVI-like federation in the future.

The NOVI IM uses a semantic web approach and it is complemented by data models that use the *Web Ontology Language* (OWL). This choice has been driven by the desire to support *reasoning* and *context awareness*, which in turn allow NOVI to create efficient and complex services with resources available within the federation. Existing IMs do not fulfill all the expected requirements and cannot support the description of shared resources and services within a federation of heterogeneous platforms. We decided nonetheless to use the ontologies and experiences from some of the other efforts to better align the NOVI model in this ecosystem. Existing models that provide direct inputs into the development of the NOVI IM are shown in Fig. 2. Among them, the Network Markup Language (NML)[9] and Network Description Language (NDL)[10] provided strong influence in specifying the NOVI *resource ontology*.

The NOVI IM consists of three distinct yet related ontologies, thus facilitating the adoption of its modules by communities interested in specific aspects. Specifically, the NOVI IM defines a *resource ontology*, a *monitoring ontology*, and a *policy ontology*.

The *resource ontology* specifies the concepts and methods to describe the resources offered by FI platforms and how they are connected together in a federated environment. This ontology provides the basis for topology and request descriptions and the terminology for describing physical nodes, virtual nodes, virtual topologies, and so on. The *monitoring ontology* extends the *resource ontology* with descriptions of the concepts and methods of monitoring operations, for example, details about monitoring tools, their relationship to resources, and types of measurements that can be gathered. Finally, the policy ontology extends the resource ontology with descriptions of the concepts and methods enabling the management and execution of policies defined within member platforms of a NOVI federation. An extended description of the IM is provided in [1].

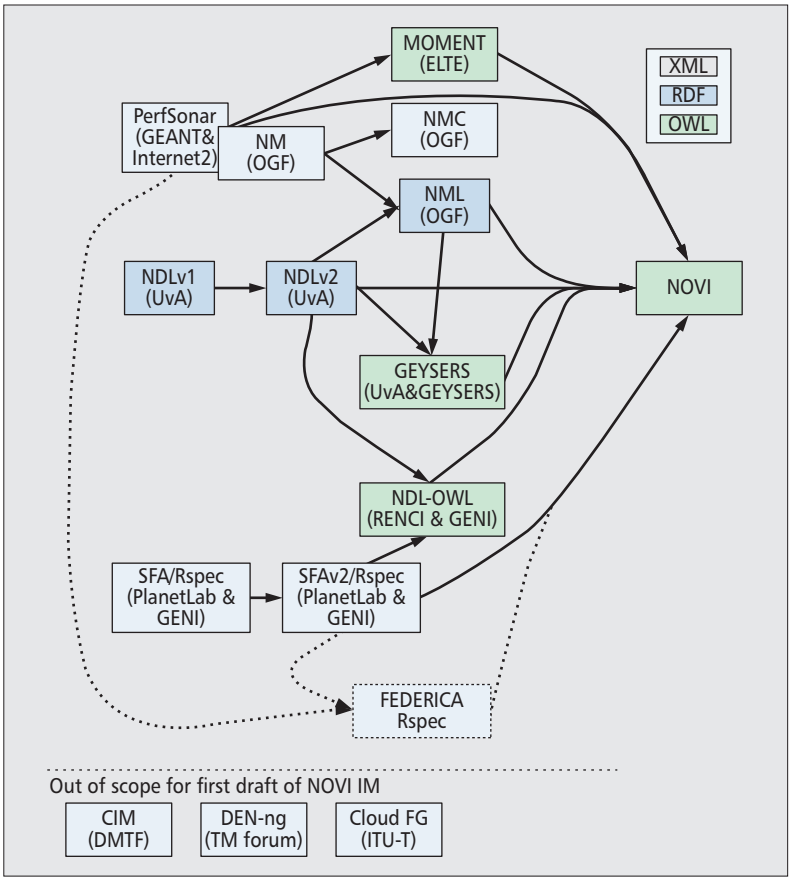


Figure 2. Relation of the NOVI information model with other IMs.

Semantic-Based Resource Discovery and Mapping Service

Resource discovery enables locating and retrieving information across the federated virtualized substrate network in a decentralized way via a scalable query process. The NOVI *resource information service* (RIS) acts as a single point of contact within the SL for other services to acquire information about the status of virtual and substrate resources. To that end, it interacts with the *request handler* to communicate with the underlying platform, to reserve resources and to obtain the resource advertisements. It uses the *monitoring service* to query on the availability and the status of the resources and the *policy service* to obtain information related to the access rights or the users.

The RIS exploits the features of the NOVI IM to improve the precision of resource discovery and to apply reasoning when selecting resources and services. The RIS copes with heterogeneity by leveraging the vocabularies provided by NOVI's IM. RIS uses a database engine based on semantic web technologies where the data is stored as RDF triples. An extended description of the RIS is provided in [2].

The *intelligent resource mapping* (IRM) service is responsible for mapping user requests for virtual topologies to the federated physical substrate. The IRM service gathers information from the RIS regarding substrate resources availability, utilization, and access control. The need for efficient sharing of virtualized infrastructures within NOVI has led to the introduction of novel techniques related to the *inter-domain VNE problem*, as reported in [2, 7, 8]. In order to solve the problem, the IRM service breaks it down to the following phases.

The VN Partitioning Phase: During this phase, user requests for VNs are split by a local instance of the IRM service into partial requests, which are apportioned to the platforms-members of the federation in a cost-efficient way. Request splitting is based on appropriately defined resource provisioning costs, as reported in [8]. To deal with the inherent complexity and scalability of the VN partitioning problem, a request partitioning algorithm with the use of iterated local search meta-heuristic has been introduced in [8].

The Intra-Domain VNE Mapping Phase: For a given VN partition, this phase provides an assignment of user VN requests to specific substrate nodes and links within a single administrative domain. In other words, VN embedding sub-problems are formulated and executed on

[9] Network Mark-up Language Working Group (NML-WG), http://www.ogf.org/gf/group_info/view.php?group=nml-wg

[10] Network Description Language (NDL), http://www.science.uva.nl/research/sne/ndl

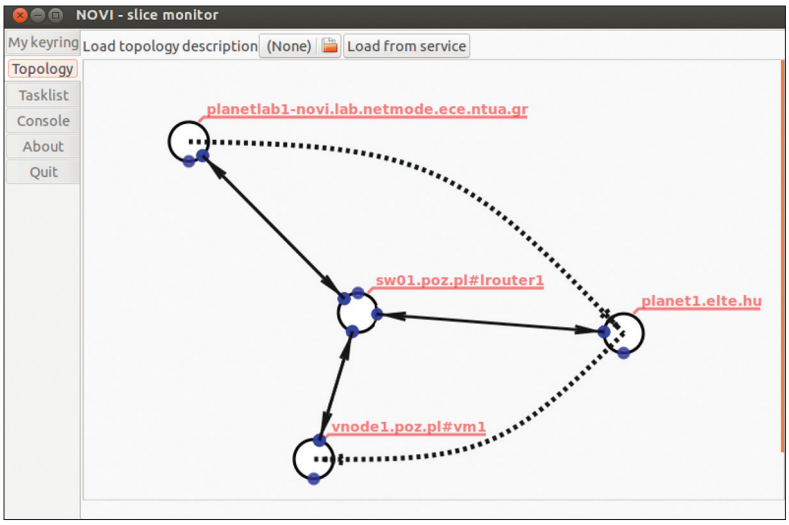


Figure 3. GUI of the NOVI monitoring service.

each IRM service of the platform-members selected at the VN partitioning phase, resulting into sub-optimal allocation of virtual resources within the federated substrate. Resource mapping within the context of NOVI depends on the characteristics of the testbed and the requested resources and constraints imposed by the user. Therefore, in order to map VN requests on the FEDERICA substrate, we followed the methodology introduced in [7] denoted as networked cloud mapping, while for PlanetLab a modified version of the semantic based approach for VN mapping presented in [2] has been adopted.

NOVI Monitoring Framework

A key component of NOVI's control and management plane is the *monitoring service* (MS), responsible for both substrate and slice monitoring. In the former, monitoring is performed on the physical substrate resources (physical hosts, links, paths, etc.), while in the latter monitoring is performed on slices composed of virtualized resources allocated to a user. The NOVI experiments use diverse monitoring tools deployed within the federated infrastructures, leading to two challenges:

- Handling the monitoring tools in a common way.
- Harmonizing the measurement data produced by the underlying tools.

To this end, we developed a generic *monitoring ontology* (MO), enabling us to describe, parameterize, and use various active and passive monitoring tools installed within the different federated platforms.

Depending on the usage scenario, the MS can support two main tasks. The first task is triggered by RIS prior to resource allocation in order to collect monitoring and measurement information from the substrate, which can be used by IRM to ensure that the constraints defined in the resource requests are satisfied. The second task is used after the resource reservation process, to perform slice monitoring for diagnostic purposes, that is, checking the current status of a given set of virtual resources across the NOVI federation. MS supports advanced monitoring tools that enable users to measure key performance metrics of the network (e.g. one-way delay, round-trip time, packet loss, available bandwidth). Testbed-specific configuration ontologies [1] are used to describe the testbed-specific implementation of metrics to be monitored, enabling MS to automatically instantiate monitoring tasks in a federated environment.

In order to help users with slice monitoring, we developed a graphical front-end. For example, the experimenter can define a resource-specific measurement during an experiment in their virtual topology as depicted in Fig. 3, by clicking on the selected resource.

Measurements of selected metrics can be managed individually, based on user-specified monitoring tasks that can be started, stopped, or removed from the task list of the GUI. Measurements can be viewed via the GUI, uploaded to a database within the resource information service, or even trigger event-condition-action policies in the policy service (see section below). Figure 4 illustrates that the user can retrieve real-time monitoring information (e.g. memory utilization) through the monitoring GUI.

NOVI Policy-Based Management System

The *Policy Service* (PS) is used in NOVI as a management service controlled by policies, following the *Policy-based Network Management* (PBNM) [9] approach and building upon the Ponder2 policy framework.[11] Using the abstractions of the NOVI IM, we define and enforce domain-independent management policies for both intra-domain and inter-domain management purposes, considering as a domain a member platform of a NOVI federation, that is, one domain could be the PlanetLab platform while another domain could be the FEDERICA platform.

Intra-domain policies are specified in the Ponder2 policy language. There is support for:

- *Access control policies* that specify which rights users have on specific resources.
- *Event-condition-action policies* enforcing management actions upon events indicating failures or performance degradation, where events are received by the MS.
- *Role-based access control* (RBAC) policies where the notion of role provides a semantic grouping of policies with a common subject, generally pertaining to a position within a platform-member of a NOVI federation, for example, user, administrator, or principal investigator in the PlanetLab experimental platform.

Inter-domain policies, defined in the form of *Ponder2 mission policies*, denote the duties of the remote domain in terms of obligations it must enforce, i.e. the management obligations that a platform must fulfill against another platform in a NOVI federation. These obligation policies are written in terms of the corresponding interfaces for each domain, denoting *events*, *notifications*, *local actions*, and *remote actions*.

Events: These refer to a specific cross-domain operation and can trigger its policies. They are the events that can be received by the corresponding interface in order to perform the actions based on inter-domain policies.

[11] Ponder2, http://ponder2.net

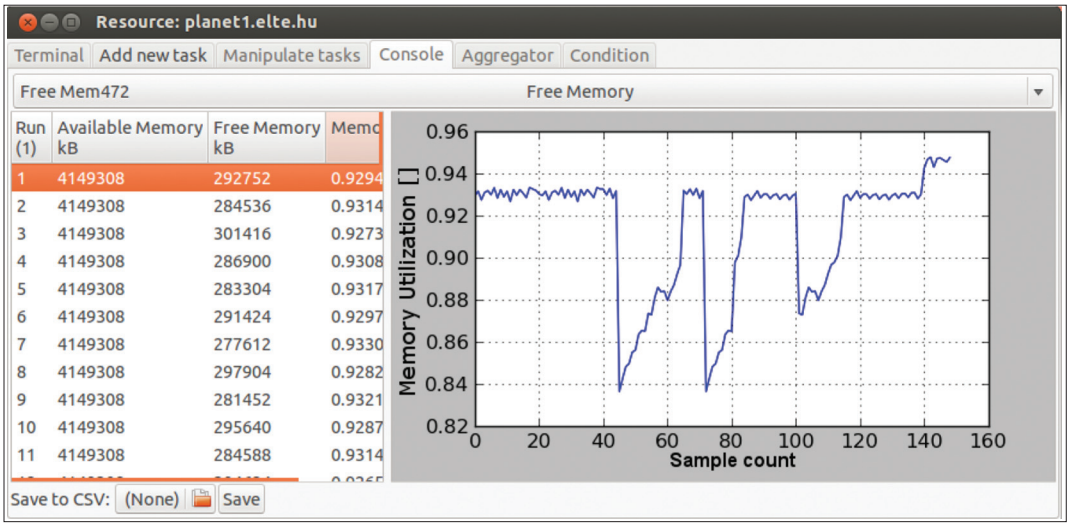


Figure 4. Retrieving real-time monitoring information through the NOVI MS GUI.

The NSwitch distributed software complements NOVI's federation architecture by providing a unified way of interaction between heterogeneous domains at the data-plane. It enables a virtual entity in one domain to be connected at protocol Layer 2 with another virtual entity in a remote domain taking into account concurrence, isolation, elasticity, and programmability aspects.

Notifications: These are events that the inter-domain policy can raise within either the local domain in which it has been loaded, or a remote domain. They can trigger actions based on other policies that exist in each platform.

Local actions: These may be invoked by the inter-domain policies in the domain in which it has been loaded. They may be management actions on local physical or virtual resources. They are expressed as methods in the managed objects that represent each resource.

Remote Actions: These may be invoked by an inter-domain policy either when it is running locally or in a remote domain. They control and manage resources in a remote platform based on the agreement between domains.

Network Stitching

Assigning and using virtual resources in a federated environment requires configuration, management, and control of multi-domain VNs across an interconnecting substrate. Although federation architectures, such as SFA, are capable of providing a homogenized way of interaction between substituent domains, they do not explicitly deal with data-plane stitching that is the prerequisite for multi-domain operations. The problem becomes more challenging when dealing with widely heterogeneous network resources (e.g. virtual machines with different hypervisors, user-defined instances within dissimilar networking equipment). This loosely coupled environment needs to be managed and configured within a stitched substrate, across diverse networking technology domains.

NOVI's network stitching approach was implemented by introducing the *NOVI Switch* (*NSwitch*) component, a software programmable switch, based on the *Open vSwitch* (*OVS*). The NSwitch distributed software complements NOVI's federation architecture by providing a unified way of interaction between heterogeneous domains at the data-plane. It enables a virtual entity in one domain (e.g. PlanetLab) to be connected at protocol Layer 2 (L2) with another virtual entity in a remote domain (e.g. FEDERICA) taking into account concurrence, isolation, elasticity, and programmability aspects. This is achieved with the use of a specific implementation of a distributed component (NSwitch driver) per testbed.

Specifically, in order to map PlanetLab slivers into an L2 broadcast domain, we adopted an approach similar to the one developed within the *VINI*[12] project in the U.S. that used an *Ethernet over GRE* (*EGRE*) mechanism to provide point-to-point virtual network capabilities to user configured virtual resources over the Internet. NOVI's NSwitch driver at the PlanetLab testbed enhanced VINI's capabilities by introducing the OVS software in PlanetLab's host operating system, thus enabling point-to-multipoint virtual links. Inside the FEDERICA domain, L2 data plane connectivity is provided by means of VLANs used by logical routers, switches, and VMs. The FEDERICA NSwitch driver relies on processes running in a border node with a public IP interface toward the PlanetLab side and a L2 connection toward the FEDERICA side. In essence, it performs the translation of EGRE key values of packets originating from PlanetLab slivers to VLAN IDs of FEDERICA resources (Fig. 5).

The distributed nature of the NSwitch mechanism does not introduce scalability limitations as far as the total number of concurrent network slices is concerned. The use of well known and widely supported network protocols (GRE tunnels and IEEE 802.1Q VLANs) ensures compatibility with a large number of FI infrastructures. Finally, the NSwitch does not introduce any significant performance degradation on end-to-end delays and bandwidth between virtual entities, compared to physical (substrate) entities [10] .

NOVI Service Layer Interfaces

The NOVI service layer (Fig. 1) supports two interfaces:

- A northbound interface (*NOVI API*) that provides the means to the user for interacting with the NOVI environment.
- A southbound interface (*request handler*) that is responsible for the communication of the NOVI service layer with the federated testbeds.

[12] VINI — A VIrtual Network Infrastructure, http://www.vini-veritas.net

NOVI provides methods and tools to federate heterogeneous FI infrastructures. Experimental testbeds can be added to and removed dynamically from the NOVI innovation cloud. These testbeds are managed by separate, yet interworking providers that can also add/remove resources in each underlying substrate dynamically.

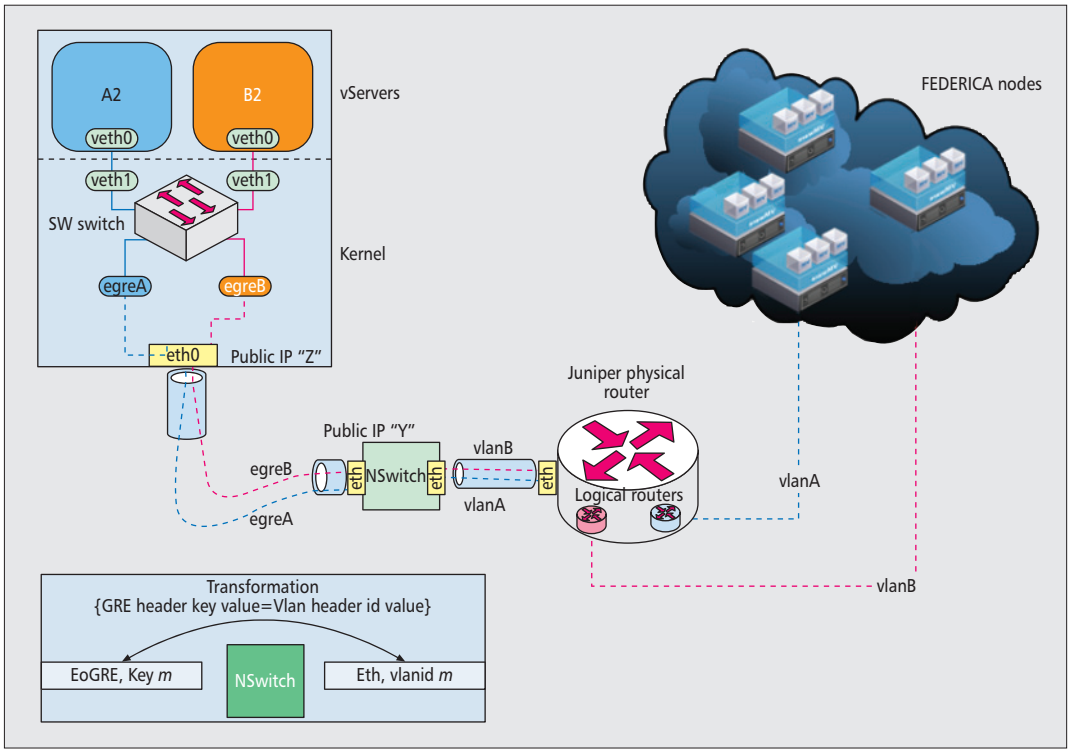


Figure 5. L2 network federation example of heterogeneous virtual infrastructures (GRE & 802.1q protocols are used for L2 topology creation).

The NOVI API: Provides the entry point for interacting with the NOVI control and management services. It has three main tasks:

- It accepts requests from authenticated users containing resource requirements represented in the NOVI IM.
- It handles and delivers the request to the appropriate component within the NOVI service layer.
- It provides feedback to the users on how their request is handled before the experiment starts being executed in a combined NOVI slice.

The user can follow the communication between NOVI components in real-time and assess the status of their request via a Web-based GUI provided on top of the NOVI API. Thus, a user can define a virtual network topology along with the characteristics for requested resources. For every request, the GUI generates an OWL document based on the NOVI IM, which is sent to the NOVI API as an HTTP request.

The Request Handler (RH): Acts as an intermediate component between NOVI control and management services and the federated substrate. Its main purpose is to perform two types of operations:

- Delivering resource allocation requests to the underlying platforms.
- Enabling the RIS to retrieve information from testbeds that are members of a NOVI federation.

To this end, the NOVI IM specifications need to be translated into platform-specific specifications. In our testbed implementation, this entailed translating NOVI IM concepts into SFA RSpec based on the ProtoGENIv2 RSpec.[13] Note that PlanetLab resource specifications are SFA enabled. However, in the case of FEDERICA, several extensions had to be made for SFA compliance of network-specific resources, for example, logical routers. Via this translation mechanism, the RH enables users with a unified method to create, update, and delete resources of his/her slice.

Scalability Issues

NOVI provides methods and tools to federate heterogeneous FI infrastructures. Experimental testbeds can be added to and removed dynamically from the NOVI innovation cloud. These testbeds are managed by separate, yet interworking providers that can also add/remove resources in each underlying substrate dynamically. The NOVI SL should be able to leverage the addition of extra resources and testbeds. Therefore, we investigated key scalability issues of the NOVI framework such as the scalability of the inter-domain topology embedding methodology used in the NOVI framework.

Large Scale Topologies: We have evaluated experiments requesting interconnected topologies of growing sets of resources per request with regards to VN partitioning. The evaluation results are reported in [8] and indicate that the implemented approach effectively addresses time performance/scalability issues for large incoming requests (less than three seconds for a partial mesh VN comprised of 350 nodes). Partitioning time increases linearly with the number of requested nodes at less than three percent increase in the partitioning cost compared to an exact algorithm.

[13] ProtoGENI, http://www.protogeni.net/trac/protogeni/wiki/RSpec

Horizontal Scalability: The impact of the addition/removal of testbeds within the NOVI federation was also evaluated in [8] with regards to the topology embedding approach. Two sets of five and 10 distinct infrastructure providers have been tested, signifying the effectiveness of the implemented approach in a multi-domain environment (less than two percent increase in the partitioning cost compared to an exact approach).

Regarding the addition/removal of resources in each underlying platform (vertical scalability), the performance of the resource allocation process highly depends on the specific resource mapping algorithm being used by each provider, thus it is outside the realm of NOVI.

Demonstrations and Usage Scenarios

The NOVI prototype was demonstrated in two flagship European Union events, namely *FIA*-2011 (Future Internet Assembly, Poznan, Poland, October 2011) and *FIA*-2012 (Future Internet Assembly, Aalborg, Denmark, May 2012). In the former, the working prototype of the NOVI framework was demonstrated, while the latter demonstrated the creation of two concurrent interconnected slices utilizing specific resources from both PlanetLab and FEDERICA platforms, as depicted in Fig. 6.

The NOVI testbed was used to study network performance aspects requiring a diversity of network characteristics (QoS constraints). For example, we conducted a video streaming experiment across FEDERICA and PlanetLab platforms. To conduct this experiment, a VLC video server was installed in a FEDERICA VM located within a FEDERICA host, while several video clients were launched in both FEDERICA VMs and PlanetLab hosts. During the execution of the experiment we obtained QoS-related statistics with the use of the VLC player software in the client side. The results showed that FEDERICA users perceived considerably better performance than PlanetLab clients, as the video was produced from a QoS-enabled environment (FEDERICA) and routed to clients in a non QoS-enabled environment (PlanetLab) over the best-effort public Internet. This behavior highlights the ability of the NOVI platform to cater to user-driven experimentation in diverse interconnected networking environments, which may arise as a requirement in federated usage scenarios.

Finally, the NOVI federation framework can be exploited by FI researchers, and it can also be used as an academic tool for graduate/undergraduate courses. As a proof of concept, academic NOVI participants (NTUA, ELTE, UvA, and UPC) created *NOVI lab* exercises that were delivered to more than 100 students in the fall 2012 semester, as part of their academic training. The lab exercises consisted of slice creation, update and deletion operations on the NOVI integrated testbed, with the use of the NOVI GUI. Instructions for three types of slices were instantiated by the students via unbound, bound, and partially bound requests.

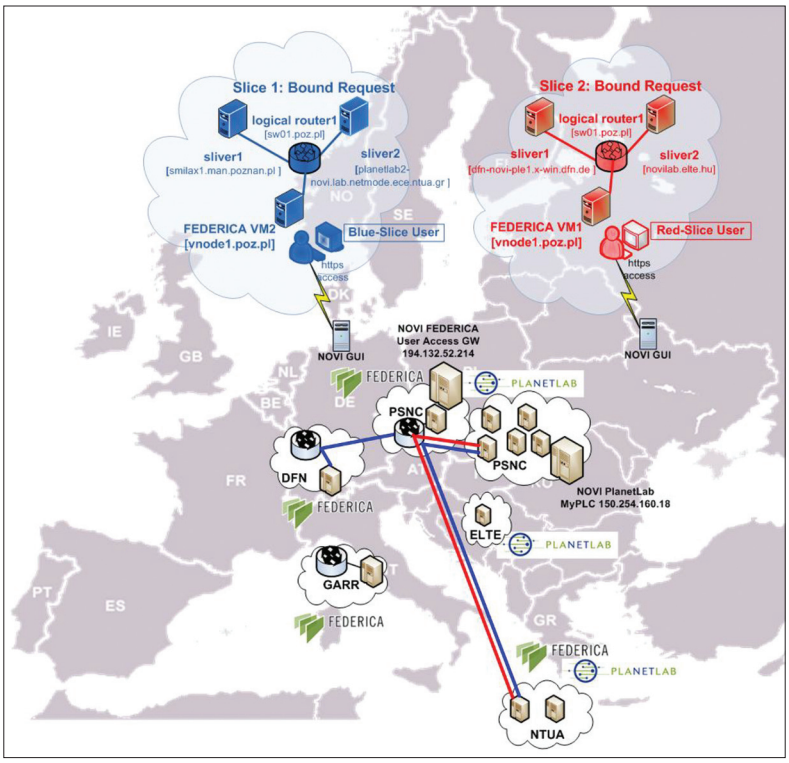


Figure 6. Instantiation of two concurrent slices in the federated NOVI environment.

Conclusion

The NOVI framework specifies a distributed data, control, and management plane federation architecture that respects individual domain mechanisms and can be applied on stitched heterogeneous FI testbeds. As a proof of concept, the NOVI framework has been developed and successfully applied to federate a private PlanetLab domain with resources interconnected over the best-effort Internet and FEDERICA, an infrastructure of virtual resources interconnected via dedicated L2 networking facilities of European NRENs and GÉANT. NOVI's prototyped architecture demonstrated the applicability and validity of the framework on extending virtualization across dissimilar platforms, resources (i.e. computing and networking), and protocols.

Further development and adoption of the NOVI platform can empower smart, user-controlled configurations of virtual resources homed in dissimilar yet stitched communities. NOVI participants are active players in FI research and innovation, both academic and industrial. Their extensive experimentation on the NOVI testbed demonstrated the capabilities and limitations of individual and federated platforms. This pre-normative work could contribute to bridging FI experimental federations with interconnected cloud architectures and interworked public/private data-centers, adding value via its intelligent services, IMs, and composite algorithms.

Acknowledgment

This work was partially supported by the European Commission, 7th Framework Programme for Research and Technological Development,

NOVI participants are active players in FI research and innovation, both academic and industrial. Their extensive experimentation on the NOVI testbed demonstrated the capabilities and limitations of individual and federated platforms.

Future Internet Research & Experimentation (FIRE), Grant no. 257867 — NOVI. Extensive documentation of NOVI services and experiments are provided in its public deliverables, accessible at *http://www.fp7-novi.eu/deliverables*

References

[1] J. van der Ham *et al.*, "The NOVI Information Models," *Future Generation Computer Systems*, vol. 42, Jan. 2015, ISSN 0167-739X, pp. 64–73.
[2] C. Pittaras *et al.*, "Resource Discovery and Allocation for Federated Virtualized Infrastructures," *Future Generation Computer Systems*, vol. 42, Jan. 2015, ISSN 0167-739X, pp. 55–63.
[3] M. Berman *et al.*, "GENI: A Federated Testbed for Innovative Network Experiments," *Elsevier Computer Networks* (special issue on Future Internet Testbeds), vol. 61, Mar. 2014, pp. 5–23.
[4] T. Rakotoarivelo, G. Jourjon, and M. Ott, "Designing and Orchestrating Reproducible Experiments on Federated Networking Testbeds," *Elsevier Computer Networks*, vol. 63, Apr. 2014, pp. 173–87.
[5] L. Peterson and T. Roscoe, "The Design Principles of PlanetLab," *ACM SIGOPS Operating Systems Review*, vol. 40, no. 1, Jan. 2006, pp. 11–16.
[6] P. Szegedi *et al.*, "With Evolution for Revolution: Managing FEDERICA for Future Internet Research," *IEEE Commun. Mag.*, vol. 47, no. 7, July 2009, pp. 34–39.
[7] C. Papagianni *et al.*, "On the Optimal Allocation of Virtual Resources in Cloud Computing Networks," *IEEE Trans. Computers*, vol. 62, no. 6, June 2013, pp. 1060–71.
[8] A. Leivadeas, C. Papagianni, and S. Papavassiliou, "Efficient Resource Mapping Framework over Networked Clouds via Iterated Local Search-Based Request Partitioning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 6, June 2013, pp. 1077–86.
[9] M. S. Sloman, "Policy Driven Management for Distributed Systems," *Journal of Network and Systems Management*, vol. 2, no. 4, 1994, ISSN:1064-7570, pp. 333–60.
[10] C. Argyropoulos *et al.*, "Network Virtualization over Heterogeneous Federated Infrastructures: Data Plane Connectivity," *IFIP/IEEE Integrated Network Management Symp.* (IEEE IM 2013), Ghent, Belgium, May 2013.

Biographies

V. Maglaris holds an engineering diploma from the National Technical University of Athens – NTUA (1974) and a Ph.D. from Columbia University (1979). Between 1979 and 1989 he held industrial and academic positions in the USA. Since 1989 he has been a professor at the School of Electrical & Computer Engineering of NTUA, teaching and performing research on Internet technologies. He served as the Chairman of the GÉANT Consortium (2004–2012) that governs the Pan-European Research and Education Network. From July 2012 to June 2013 he was the General Secretary for Research & Technology of the Greek Ministry of Education.

C. Papagianni received her diploma and Ph.D. in electrical and computer engineering, both from NTUA, in 2003 and 2009, respectively. Since 2010 she has been a senior R&D associate at the Network Management & Optimal Design (NETMODE) Laboratory of NTUA. Her research interests are in the area of computer networks with an emphasis on cloud computing, virtualization, network optimization and management, and service provisioning.

Georgios Androulidakis (gandr@netmode.ntua.gr) received his Ph.D. and diploma in electrical and computer engineering from the National Technical University of Athens (NTUA), Greece in 2009 and 2003, respectively. His research interests are in the area of computer networks with an emphasis on network security, traffic analysis, network virtualization, and software defined networking.

M. Grammatikou was born in Athens, Greece. She holds a diploma and a Ph.D. in electrical and computer engineering, both from NTUA. She has been a senior researcher at the NETMODE laboratory of NTUA since 2001, teaching and performing research on computer networks and cloud computing.

P. Grosso is an assistant professor in the SNE group at the UvA. She is the lead researcher of the group's activities in the fields of distributed infrastructure information modelling, SDNs, and Green ICT. She participates in several national and international projects, such as ENVRI-PlUS, GN4, COMMIT, GreenClouds.

J. van der Ham received his Ph.D. in 2010 at the University of Amsterdam, titled "Semantic descriptions of complex computer networks". He was the editor of the NML schema document, and lead designer of the NOVI-IM.

C. de Laat chairs the System and Network Engineering (SNE) research group at the University of Amsterdam. Research ranges from advanced networking for processing of big data in PetaScale e-Science applications, semantic web to describe e-infrastructure resources, information complexity, authorization architectures, and systems security and privacy of information in distributed environments.

B. Pietrzak received the M.Sc. degree in computer science on the software engineering faculty from the Poznan University of Technology in 2003. He joined the Poznan Supercomputing and Networking Center (Poland) as a senior network engineer in 2004. He worked in a number of senior management roles in a range of technology driven projects, including NOVI, IIP, and GN3plus.

B. Belter received the M.Sc. degree in computer science from the Poznan University of Technology in 2002. He works at the Poznan Supercomputing and Networking Centre as Head of the Next Generation Networks Department. He participated in several FP6/FP7 projects, including 6NET, PHOSPHORUS, GEANT, GEYSERS, BonFIRE, and NOVI. He also participated in a number of national initiatives funded by the Polish Ministry of Science and Higher Education. Currently he is coordinating two research projects: H2020 COMPLETE and FP7 FELIX. He was a member of the TERENA Networking Conference Programme Committee. His main research activities concern the architectural aspects of control and management planes in optical networks, and quality of service in next generation packet networks. He is the author or co-author of papers in professional journals and conference proceedings.

J. Steger was born in Székesfehérvár, Hungary in 1976. He graduated as a biophysicist at Loránd Eötvös University, Budapest, Hungary (2001), and earned the Ph.D. Degree in Physics (2010). He works as a junior assistant professor. His teaching and scientific interests include signal processing, measuring and modelling communication networks, and complex system simulation techniques. He has worked for several domestic and EU projects, and is a member of the Communication Network Laboratory, Budapest, Hungary.

S. Laki received the M.Sc. and Ph.D. degrees in computer science from the Eötvös Loránd University, Budapest, Hungary, in 2007 and 2015, respectively. He is currently working as an assistant professor in the Department of Information Systems, Eötvös University. His research interests include active and passive network measurement techniques, traffic analytics and algorithmic aspects of peer-to-peer and other networks.

M. Campanella started working on computers and networks in 1984. He is currently acting as research coordinator for the Italian research and education network (GARR). He works in various roles in European projects on networking, including GÉANT, coordinated the FEDERICA project on future Internet, and is currently working on SDN.

S. Sallent received an M.Sc. degree (1979) and a Ph.D. degree (1988) in telecommunications engineering, both from the Technical University of Catalonia (UPC), in Barcelona, Spain. His research interests include resource optimization and scheduling, optical access communications, Internet architectures, and traffic modeling. From 1979 to 1985 he was with Philips Company. Currently he holds the position of full professor at UPC, where he leads the Broadband Networks research group within the Department of Network Engineering. He is also the Director of the i2Cat Foundation, in Cataluña, Spain. He was President of the Spanish Telematic Association. He has participated in several research projects, funded by the EU (FEDERICA, Phosphorus, NOVI, Euro-NF, and FIBRE), the Spanish government, and private companies.

Quality of Experience Management in Mobile Cellular Networks: Key Issues and Design Challenges

Eirini Liotou, Dimitris Tsolkas, Nikos Passas, and Lazaros Merakos

ABSTRACT

Telecom operators have recently faced the need for a radical shift from technical quality requirements to customer experience guarantees. This trend has emerged due to the constantly increasing amount of mobile devices and applications and the explosion of overall traffic demand, forming a new era: "the rise of the consumer". New terms have been coined in order to quantify, manage, and improve the experienced user quality, with QoE being the most dominant one. However, QoE has always been an afterthought for network providers, and thus numerous research questions need to be answered prior to a shift from conventional network-centric paradigms to more user-centric approaches. To this end, the scope of this article is to provide insights on the issue of network-level QoE management, identifying the open issues and prerequisites toward acquiring QoE awareness and enabling QoE support in mobile cellular networks. A conceptual framework for achieving end-to-end QoE provisioning is proposed and described in detail in terms of its design, its constituents and their interactions, as well as the key implementation challenges. An evaluation study serves as a proof of concept for this framework, and demonstrates the potential benefits of implementing such a quality management scheme on top of current or future generations of mobile cellular networks.

INTRODUCTION

Quality of Experience (QoE) is defined by ITU-T as *"the overall acceptability of an application or service, as perceived subjectively by the end-user"*. Otherwise put, it describes the degree of the end-user's *"delight or annoyance"* when using a product or service [1]. Inherently, QoE is a very broad and generic concept, and as such it incorporates any conscious or unconscious aspects that affect overall user satisfaction.

This generic notion of QoE has opened up research to a variety of systems and application domains. In this study, we narrow down the scope to the telecommunications domain, where QoE intelligence is of crucial importance, not only to the end-consumers but also any stakeholders involved in the service provisioning chain. In telecommunication networks, despite the catholic presence of inherently deployed Quality of Service (QoS) mechanisms, QoE has been an "afterthought". No generation of telecommunication networks has been originally designed with QoE principles so far. Nevertheless, the system-centric view of QoS provisioning is no longer sufficient, and it needs to be replaced or complemented with more user-centric approaches [2]. Therefore, the shift from QoS-centric to QoE-centric networks is an emerging, open challenge.

Toward this direction, new architectures have been proposed regarding the collection and exploitation of QoE-related information. For instance, a block diagram for the QoE management of next generation networks (NGNs) is proposed in [3], where adaptations to the NGN-specific network attachment control and resource and admission control functions are described. Furthermore, a novel architecture for QoE support in Long Term Evolution (LTE) systems requiring new, proprietary interfaces is described in [4]. Other works focus on specific services, such as the customer experience management (CEM) system for IPTV described in [5]. Similarly to the aforementioned examples, the majority of current works proposes solutions tailored to concrete systems or services. In parallel, standardization activities mainly handle the issue of QoE estimation, a.k.a. "QoE modeling" [6], leaving the end-to-end QoE provisioning realization out of the discussion. Motivated by this observation, this article proposes the required steps for enabling QoE-based management in the environment of mobile cellular networks. Our contribution lies in identifying the design challenges and requirements toward QoE provisioning, namely a) gaining QoE awareness and b) using this awareness to enable effective QoE-centric decisions on top of mobile cellular networks (e.g. GSM, UMTS, LTE/LTE-A). In this way, a better understanding of the challenging

The authors are with National and Kapodistrian University of Athens.

0163-6804/15/$25.00

On the one hand, this heterogeneity better supports the ever increasing traffic requirements, pushing toward an increase of the end-user QoE, while on the other hand it imposes higher interference and more severe competition over the scarce spectrum resources, pushing toward a QoE decrease.

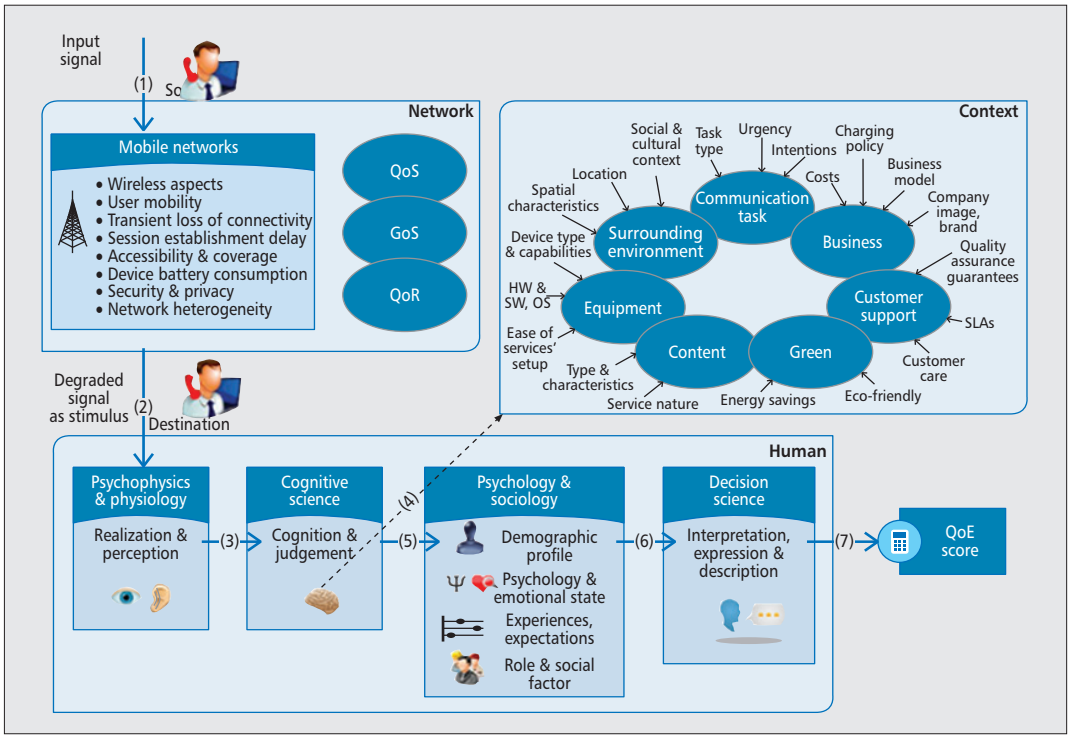


Figure 1. The three QoE influence pillars.

topic of QoE in mobile cellular networks is gained.

The remainder of this article is organized as follows. We first provide a comprehensive composition of the QoE notion from a mobile cellular network's perspective by describing, in an end-to-end manner, the most important factors influencing quality. Then we present a conceptual framework toward QoE support, described in terms of functionalities, interactions, and design challenges. Afterward, realization issues for the tight integration of the proposed QoE provisioning framework in mobile cellular networks are identified, and evaluation results are presented, using the LTE network as a case study.

The Ingredients of QoE

QoE is a broad concept, embracing influence factors from different domains and disciplines. We adopt the approach of [1] and categorize those factors into three major pillars, namely *system* (here, *network*), *human*, and *context*, which taken together, formulate the overall user QoE. Moving one step further, we group the most dominant factors per pillar, and illustrate how QoE opinions are progressively formed during a communication session (i.e. how these pillars are connected) (Fig. 1).

The *network* pillar consists of any end-to-end quality affecting parameters, as these are described by the Quality of Service (QoS), Grade of Service (GoS), and Quality of Resilience (QoR) terms [2]. It embraces technical characteristics of the traversed network, equipment specifications, application characteristics, and so on. This pillar is strongly connected with network-specific factors, which are particularly important and decisive for the operator (see the "Network" box in Fig. 1). In the case of mobile cellular networks, the most challenging and less investigated factor is their inherent heterogeneity, referring to the dynamic emergence of geographically distributed and overlapping smaller cells. On the one hand, this heterogeneity better supports the ever increasing traffic requirements, pushing toward an increase of the end-user QoE, while on the other hand it imposes higher interference and more severe competition over the scarce spectrum resources, pushing toward a QoE decrease.

Moving on to the *human* pillar, we describe it as the superset of four subcategories, where each one comprises a unique scientific area that influences the overall user's quality impression. Initially, the area of Psychophysics quantifies the relationship between a physical stimulus (e.g. sound/image) and the resulting perception to the human sensory system. Then, Cognitive Science studies the human mind and how this works in terms of interpretation, reasoning, judgment, information processing, and so on. Psychology and Sociology help understand the human character and behavior both as a unity and as part of the society, which uniquely affects the user's understanding of quality. Finally, Decision Theory studies the rationality and optimality in decision making.

Finally, the *context* includes any kind of background information that consciously or unconsciously affects the user's judgment. For instance, QoE is influenced by the spatiotemporal environment where the service is provisioned (open-air crowded place vs. quiet office); the equipment under use (mobile phone vs. tablet); the service and content type (audio/video/text/graphics); the content characteristics (head-and-shoulder video vs. football game); the communication task (public safety vs. leisure browsing); and other contextual information related to busi-

ness or financial aspects (e.g. charging policy, marketing, brand effect).

Depicted in Fig. 1 is the progressive formulation of an end-user's QoE during a communication session, presented in chronological order (steps (1)–(7)). One source-generated signal is entering the network (1), and its distorted version reaches its destination (2), where it is perceived by the target end-user as a visual/audio stimulus (3). This stimulus is internally represented into the human brain, processed as information content and in terms of quality (5). This quality judgment is significantly affected by numerous external factors, which all together constitute the context of this communication scenario ((4)-dashed). Following this, the quality impression is further influenced by unique characteristics of the human subject (e.g. demographic profile, current psychology, expectations) (6). Finally, the formed quality perception is expressed as a QoE score in a given scale (7), such as the five-point mean opinion score (MOS).

A Conceptual Framework Toward QoE Management in Mobile Cellular Networks

In this section we propose a framework that enables QoE management in mobile cellular networks. To this end, we identify its required building blocks, their inner functionalities, and in-between interactions.

The structure of the quality provisioning chain and the required interactions ((1)–(6)) are presented in Fig. 2. In the core of the proposed framework is a *central QoE management entity*, which is implemented at an administrative location of the operator's network, on top of the mobile cellular network depicted in Fig. 2 by the "Network" cloud. This entity is able to collect QoE-related input and apply QoE-driven network management decisions. It consists of three main building blocks: the *QoE-controller*, *QoE-monitor*, and *QoE-manager*.

The *QoE-controller* plays the role of an interfacing between the central entity and the underlying network, synchronizing communication exchange in both directions. It is in charge of configuring the data acquisition process, by requesting and collecting feedback from appropriate data sources (e.g. some QoS indicators), as will be further analyzed below (interactions (1) and (2) in Fig. 2, respectively). The QoE-controller also decides and imposes the periodicity of this process (through (1)), namely it controls how often QoE input should be generated/gathered, and consequently how often QoE will be assessed. Having collected the required data, this component provides input of interest both to the QoE-monitor and the QoE-manager ((3a) and (3b), respectively). More specifically, it provides QoE-input data on a per flow basis to the former, and information regarding the current network state to the latter (e.g. network topology, resource availability, etc.). Finally, the QoE-controller applies any QoE-aware control decisions back to the network, during the final step of the QoE management loop (6).

Second, the *QoE-monitor* is responsible for estimating the QoE per flow, that is, per user's session, and for reporting this to the QoE-manager (4). Using network-derived input available through the QoE-controller, the QoE-monitor initially performs traffic classification to deduce the type of traffic of the considered flow. This procedure is feasible using statistical analysis (e.g. [7]). Inside the QoE-monitor, already built-in QoE assessment functions, referred to as "QoE models" (i.e. formulas for quantifying a service's QoE) are available, different per traffic/service type (e.g. video/voice/data). Depending on the identified traffic type, the proper QoE estimation model is selected by the QoE-monitor, followed by an estimation of the QoE. It needs to be noted that all available QoE models are integrated offline into the QoE-monitor by the operators, namely during the design phase of the central QoE management entity and prior to its real-time operation, which makes the original selection of QoE models very crucial.

The last component of the proposed framework is the *QoE-manager*, responsible for conducting any type of customer experience management (5a) or QoE-aware network management (5b). It uses a) input from the QoE-controller regarding the current network state, b) estimated QoE scores through the QoE-monitor, and c) operator-specific information, such as network policies or service level agreements (SLAs), as a way to decide and dictate the necessary measures that need to be imposed on the network for solving quality problems at hand. Decisions are taken per flow or catholically, respecting user policies (e.g. subscription profile, charging information, etc.) and current network constraints (e.g. availability in resources). Any QoE-triggered decisions are clearly system-specific, in the sense that their actualization depends on the

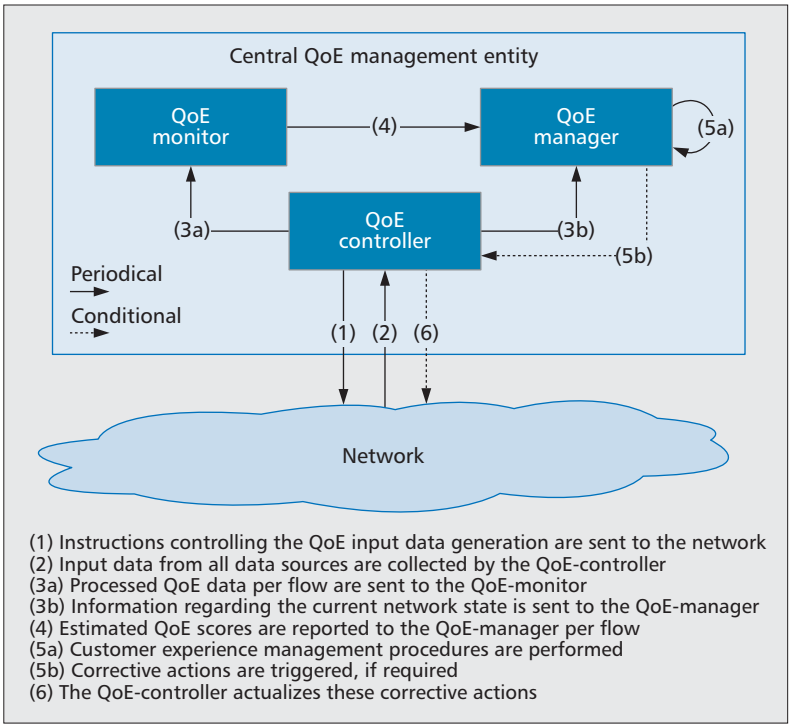


Figure 2. The proposed QoE management framework.

The QoE-controller realizes the interface between the central QoE management entity and the underlying network, by enabling a bi-directional communication exchange. Specific design decisions need to be taken when designing this building block.

underlying network. The adaptation/control actions that realize these decisions are applied to the network through the QoE-controller (6).

Next we analyze key design issues per building block, starting with the QoE-monitor, which performs the key process of estimating the QoE per flow.

The QoE-Monitor

The main challenge in the implementation of the QoE-monitor is the thoughtful selection of QoE estimation models, different per traffic/service type, to be integrated offline (a priori) into this block. QoE models imitate the human processes that occur inside a specific context each time, given the network characteristics at hand. Formally defined, a QoE model is *"a procedure that aims to model the relationship between different measurable QoE influence factors and quantifiable QoE dimensions for a given service scenario"* [8]. Consequently, the main purpose of this block is to reliably estimate QoE, as if this assessment was done by humans.

A plethora of QoE models can be found in standardization bodies' recommendations and in the literature. For instance, ITU standardization activities for IPTV QoE assessment can be found in [9], while a detailed taxonomy of objective speech quality models can be found in [10]. For VoIP services, the "E-model" is commonly used, mainly due to its valuable characteristic of providing distinct formulas for quantifying the impact of packet delays and loss rates on QoE (the "delay impairment factor" and "equipment impairment factor", respectively). For web browsing services, QoE is strongly affected by the web pages' response/loading time, and for file download services by the effective data rate. The experienced quality in real-time video applications (e.g. IPTV) is mainly influenced by the packet loss rate and burstiness, frame-rate, bitrate, and content type. Finally, the QoE for lossless video streaming services (e.g. YouTube) is significantly affected by the number and duration of stallings, as well as the video start-up delays.

QoE models are mainly classified based on their evaluation method [11]:

- *Media-layer models* make use of transmitted and/or received signals. Based on the need or not for the original source signal to be used as input, they are further characterized as full-reference, reduced-reference, or no-reference.
- *Packet-layer models* extract information from packet headers, while bitstream models use both headers and payload information.
- *Parametric models* use specific network planning parameters and metrics, as well as terminal design parameters.
- *QoS-to-QoE mapping models* are based on the non-linear dependencies between QoS parameters and QoE values.

Focusing on the challenging issue of model selection, in Table 1 representative models per evaluation method are described in terms of their applicability to mobile cellular networks (last column). To elaborate on this, information on each model's principle, required input, and purpose is also provided.

As extracted from Table 1, the use of media-layer models is not recommended for quality estimation in these networks, due to the complexity or even impossibility of setting them up. On the contrary, parametric or packet-layer models enable the acquisition of already available information through various network nodes. However, packet-layer models are not well-standardized yet, and the collection and exploitation of packet header information requires a lot of original work. Thus, presently, parametric models seem to be the perfect candidates for real-time QoE management. In addition, they require less overhead and are capable of monitoring communication sessions through heterogeneous transport infrastructures, which is ideal for modern mobile environments [10].

Another guideline for proper model selection is the avoidance of non-intrusive models, to avoid injecting pilots into the system just for QoE testing purposes and waste resources for that matter. It is preferred to exploit information already available inside the network under regular operation, which refers to actual, realistic communication scenarios. Finally, a general guideline is that the selected models for feeding the QoE-monitor block are of low complexity, well-standardized, and able to be implemented in real-time on top of existing network infrastructures.

As a final remark, it is worth mentioning, that the selective generalization of the previous model selection guidelines to other network types, such as WiFi or Ethernet, is not excluded; however, their analysis is not currently in scope.

The QoE-Controller

The QoE-controller realizes the interface between the central QoE management entity and the underlying network, by enabling a bi-directional communication exchange. Specific design decisions need to be taken when designing this building block.

Regarding the communication direction from the network to the QoE-controller (illustrated as (2) in Fig. 2), the strategic selection of appropriate nodes used for the acquisition of QoE-related input is a challenging issue. Input can be collected by various distributed nodes located at the core and access network (macro-cell/small-cell base stations, routers/servers/gateways) capturing service degradations, as well as by agents installed locally at end-devices, capturing more subjective QoE influence factors, such as context and human characteristics (Fig. 3). Some guidelines on QoE/QoS data collection in 4G networks are given in [12]. In this work, the authors propose the integration of active probes within multiple network elements between the service provider's gateway and the access network, for measuring network QoS indicators (e.g. throughput, delay, jitter), transport key performance indicators (KPIs) (e.g. round-trip times), and application/service key quality indicators (KQIs) (e.g. video frame rate, blurriness).

The appropriate type of collected QoE-related input is another important issue. This input refers to any kind of raw network data, real-time measurements, statistical/historical information, or information at the operator's possession, obtainable through:

Model type	Principle/technique	Input	Purpose/usage	Applicability to mobile cellular networks (–cons, +pros)
Media-layer: ITU-T P.862 (PESQ)	Compares the original reference signal with the degraded output signal (full reference type)	• The output (degraded) speech signal • The input (original) speech signal • Perceptual and cognitive models are required to process the two signals	•Laboratory monitoring •Live testing of prototype or emulated networks • Quality benchmarking • Codec evaluation and selection	– The reference signal is not readily available in real time => impossible for real network monitoring – Computationally heavy – Raises privacy issues – Handles the network as a "black-box" + May be used for network planning purposes
Media-layer: ITU-T P.563	Requires a preprocessing stage, a distortion estimation stage and a subsequent perceptual mapping stage	• Output speech signal • Types of calculated signal parameters are: basic speech descriptors, vocal tract analysis, speech statistics, static SNR, segmental SNR, interruptions/mutes	Non-intrusive speech quality assessment, live network monitoring and testing with unknown speech sources at the far end side	– The model may not be accurate when part of the output signal is missing due to network errors – Listening-quality evaluation only + Live end-to-end quality monitoring is feasible + Not necessarily restricted to narrowband applications
Parametric planning: ITU-T G.107	Provides a well-defined computational formula with specific input parameters	• Quality parameters from various network constituents and the terminals, e.g.: SNR, delay, equipment impairment (codec), echo, loudness, packet losses • Advantage (compensation) factor	•Performance evaluation •Network planning application and terminal design • Helps avoid "over- and under-engineering" phenomena	– In its original format the scope of the model is not for in-service evaluation – Extra signaling is required to gather and transport the model's input parameters to the QoE evaluation point – Accurate only for specific recommended application scenarios + Can be simplified to transport-level metrics for simple real-time monitoring + The analytical model formulas are not very difficult to implement
Hybrid: PSQA	Implements statistical learning tools from the random neural network (RNN) area	Network-wide and user equipment, e.g.: code, error correction, offset, packet lossrate, mean loss burst size, packetization interval, one-way delay, jitter, etc.	To dynamically optimize quality by understanding the relations among quality-affecting parameters and perceived QoE	– The model's real-time applicability is ambiguous due to human involvement for training and validation – The implementation of RNNs in the network invokes extra complexity + Can capture the human-context-technical factors' influence
Packet-layer: ITU-T P.564	Extracts information from the packets travelling in the network using DPI techniques	• Available at any mid-network monitoring points (probes, etc.), e.g.: time-stamps, sequence numbers from packet headers, and/or payload • Extra information about e.g., endpoints transferred through control packets	• Non-intrusive (passive) quality monitoring • Live assessment for reactive QoE control • SLA support	– Models have to be designed based on this Rec. (not readily available yet) – Testing is required using a detailed conformance test methodology – Extra processing load for the network midpoints to analyze packets (big data volume is produced) + Help diagnose network problems at any point + Light model
QoS-to-QoE: IQX and WQL hypotheses	• IQX: translates generic network-level QoS to QoE (exponential dependency) • WQL: describes the effect of physical stimuli on human perception-based on psychophysics (logarithmic)	Any single QoS-degrading parameter (one per mapping function), e.g., packet loss rate, setup time of a wireless connection, waiting and response times, etc.	• QoE monitoring and proactive QoE control • Already considered for video on demand quality performance assessment (e.g. YouTube)	– The mapping function has to be estimated beforehand and regularly validated – Depend on QoS parameters' availability, for IQX: network-level, WQL: application-level parameters + Enable real-time QoE control mechanisms that build on QoS monitoring + Enable smarter resource control (allocate resources when users will really perceive the difference)

Table 1. Overview of characteristic QoE estimation models.

The first possibility enabled by the proposed QoE-manager is to record and monitor real-time quality estimations per session. Acquired QoE intelligence can assist operators in comprehending and better managing their customers' overall long-term experience, increasing thereafter their loyalty level.

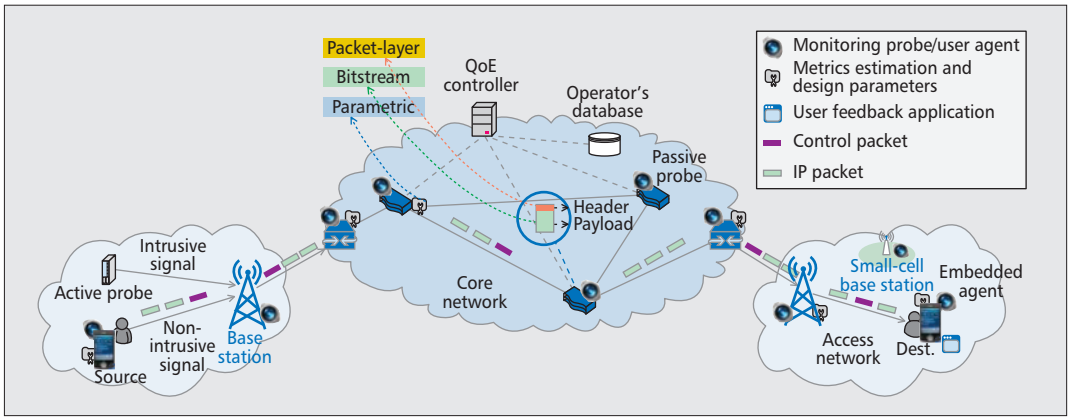


Figure 3. Illustration of the interactions between the QoE-controller and the various QoE data sources.

- Active (intrusive) or passive (non-intrusive) probes on distributed network elements.
- Embedded agents/sensors on user-devices that explicitly/silently collect usage data and statistics (e.g. monitor video playout buffers to predict stalling events).
- User-devices' applications that request user feedback.
- Any subscriber-related databases owned by the operator (Fig. 3).

The data acquisition process needs to be aligned with the QoE estimation models embedded inside the QoE-monitor. Different input parameters are required per model, and therefore the two phenomenally different procedures of the QoE-controller and QoE-monitor have to be tuned offline. Therefore, the operator's first task is to select the appropriate QoE models, and then to fine-tune the data acquisition process accordingly (Fig. 3). The collection of input may be based on packet-level information acquired through deep packet inspection (DPI) techniques (applies to packet-layer/bitstream models) or by estimating communication-related metrics (parametric models). In the case that packet-layer models are used, the characteristics and configuration of endpoints should be known in advance, or be acquired using Real-Time Control Protocol-Extensive Reports (RTCP-XR). In addition, the data acquisition procedure needs to be tuned a priori with respect to the pool of decisions/actions embedded inside the QoE-manager.

Regarding the communication direction from the QoE-controller to the network (illustrated as (1) in Fig. 2), we envision that the QoE-controller is able to dynamically configure/administrate the data generation and the data collection periodicity, for example, by switching ON/OFF some probes, based on the current network state. This periodicity needs to balance between the inevitable extra signaling overhead imposed in the network and the timeliness of the acquired data, or equivalently the accuracy of QoE estimations.

The QoE-Manager

Currently, the only opportunity for network providers to assess the offered QoE of their products or services is during the design phase, namely, prior to real-time operation. This may be accomplished by purchasing special equipment from third-party vendors, capable of performing measurements of voice/audio-visual quality through emulating the human perception. Operators may use such quality-measurement suites as a way of testing the performance of new services/devices, and thus accelerate the time-to-market. This, however, is the only course of action currently feasible; on the contrary, the proposed framework opens up possibilities for real-time quality monitoring and smart network-centric QoE management based on the operator's actual customer portfolio and realistic communication conditions.

The first possibility enabled by the proposed QoE-manager is to record and monitor real-time quality estimations per session. Acquired QoE intelligence can assist operators in comprehending and better managing their customers' overall long-term experience, increasing thereafter their loyalty level. Operators may also benefit by offering personalized services based on customer profile analytics. Moreover, the opportunity emerges for creating new QoE-based business models, to the benefit of both the users (e.g. receive differentiated quality upon demand) and the network providers (correlate charges).

Another possibility is to improve the QoE of a current flow, or to maximize the sum/average QoE of the served users catholically, for example, by expressing the total QoE as a utility function. A quality improvement may be requested either proactively or reactively. The former approach requires the prediction of network problems via QoE-based alarms, while the latter means reacting to problems already present. Potentially, any network control measures (e.g. admission control, flow prioritization, cross-layer scheduling) may be implemented, respecting network policies and constraints. The QoE-manager can also keep track of the effectiveness of these decisions, and hence be able to self-adapt and optimize the methods used for solving quality problems.

Finally, through the QoE-manager, the opportunity emerges to exploit QoE awareness as a way to potentially save on network resources without compromising the customer experience. This may become possible either by identifying moments and cases of operation when providing extra resources to an end-user would not improve the QoE perceived (e.g. [13]), or by

exploiting the non-linear relationships between QoS and QoE, such as the ones quantified by the IQX hypothesis and the WFL law [14]. The former relationship claims a negative exponential dependency between the perceived QoE values and degrading QoS parameters, while the latter describes the logarithmic impact of physical stimuli on human perception.

Enabling End-to-End QoE Support in Mobile Cellular Networks: Realization Issues and Challenges

Mobile cellular networks with QoE management aspirations may adopt and customize the proposed framework. The network-specific decisions that need to be taken are:

The physical location of the QoE management framework inside the operator's infrastructure: Challenges include determining whether this framework will be implemented as a stand-alone entity or not, centrally or in a distributed fashion, as well as developing new interfaces to support communication with other network nodes and the end-users.

The identification of the required QoE data sources, the configuration of the data collection periodicity, as well as the signaling between the network and the QoE-controller: The main concern is the minimization of the extra signaling overhead imposed on the network, compromising between scalability and estimation accuracy issues. Also, the consumed power required for QoE-data collection should be considered, mainly to avoid drainage of the handheld devices' battery.

The selection of appropriate QoE models for the QoE-monitor: Research is needed on finding ways to limit the imposed signaling required by these models, and to reduce the complexity of the QoE estimation process. Moreover, new models will need to be devised in the future, mainly to capture the long-term QoE and customer churn, based on multiple, sequential episodes with the same service. Additionally, context factors, as described in Fig. 1, need to be captured and reflected in future QoE models. Finally, traffic/service classification performed in the QoE-monitor is a very challenging issue, especially in the content-encrypted domain (e.g. HTTPS).

The type of decisions taken by the QoE-manager and their actualization through the QoE-controller: Since these decisions need to be performed on a per flow basis, and since the number of users in the network may be large, scalability and complexity issues are raised here as well.

Except for these technical challenges, the operator needs to also account for some business and legal aspects. First, ensuring end-to-end QoE may depend on multiple network, service, or content providers, especially at infrastructure inter-connection points. Therefore, collaborations and SLAs among different stakeholders are required. Second, security and privacy issues are raised, since potentially user-sensitive information has to be traversed through the network for QoE management purposes. Net neutrality issues also emerge, especially if packet differentiation is selected for QoE provisioning. Finally, the operator needs to come up with proper business cases and monetary incentives, before being convinced to implement and commercialize such a QoE management scheme.

Parameter	Values
Macro-cell radius	1 km
eNB TX power	43 dBm
HeNBs TX power	23 dBm
Number of UEs	Scalable
Distribution of UEs	Uniform inside the attached cell
Traffic load per user	1 VoIP call
VoIP codec	G.729a
Duplex mode	FDD (focus on downlink)
Channel bandwidth	10 MHz (split between macro cell and small cell)
Scheduling algorithm	Proportional fair
Flow duration	10 sec
QoE reporting period	0.1 – 10 sec
Maximum acceptable delay	0.1 sec
Packet loss robustness factor	Zero
QoE estimation model	ITU-T G.107 (E-model)

Table 2. Basic simulation input parameters.

Evaluation Results: The LTE Case Study

In this section we use LTE as a case study to demonstrate the feasibility, performance, and potential benefits of the proposed QoE management framework, using simulation. To this end, we have expanded the LTE-Sim [15] to support this framework.

We first estimate the amount of extra signaling imposed for QoE monitoring during the real-time operation of this framework, as well as the resulting accuracy of the QoE estimations. Overhead occurs due to the communication exchange between the QoE-controller and the network, whereas communication among the three main building blocks of the framework takes place internally inside the central entity. The QoE-controller is responsible for configuring the periodicity of the QoE-related data collection, referred to as the "QoE reporting period".

For this study we simulate a heterogeneous network, consisting of one macro-cell served by an evolved-eNodeB (eNB), small-cells served by home-eNBs (HeNBs) located inside 5 x 5 3GPP-based building blocks, and finally uniformly distributed user equipment (UEs). We count the

number of messages collected by the QoE-controller during configurable QoE reporting intervals (in this case, one message per UE per interval) roughly quantifying in this way the imposed overhead. With the input parameters of Table 2, we estimate how accurate the predicted QoE scores are per reporting period, using as reference the case where QoE-input is collected per 0.1 seconds. We report the obtained results in Fig. 4a, reaching the conclusion that there exists a trade-off between the amount of signaling overhead and the achieved accuracy in the QoE predictions. The results are closely dependent on the actual QoE estimation model used (here, the "E-model"), and different signaling requirements are expected by different models.

Nevertheless, this overhead may be counterbalanced considering the new opportunities enabled for QoE-driven network management.

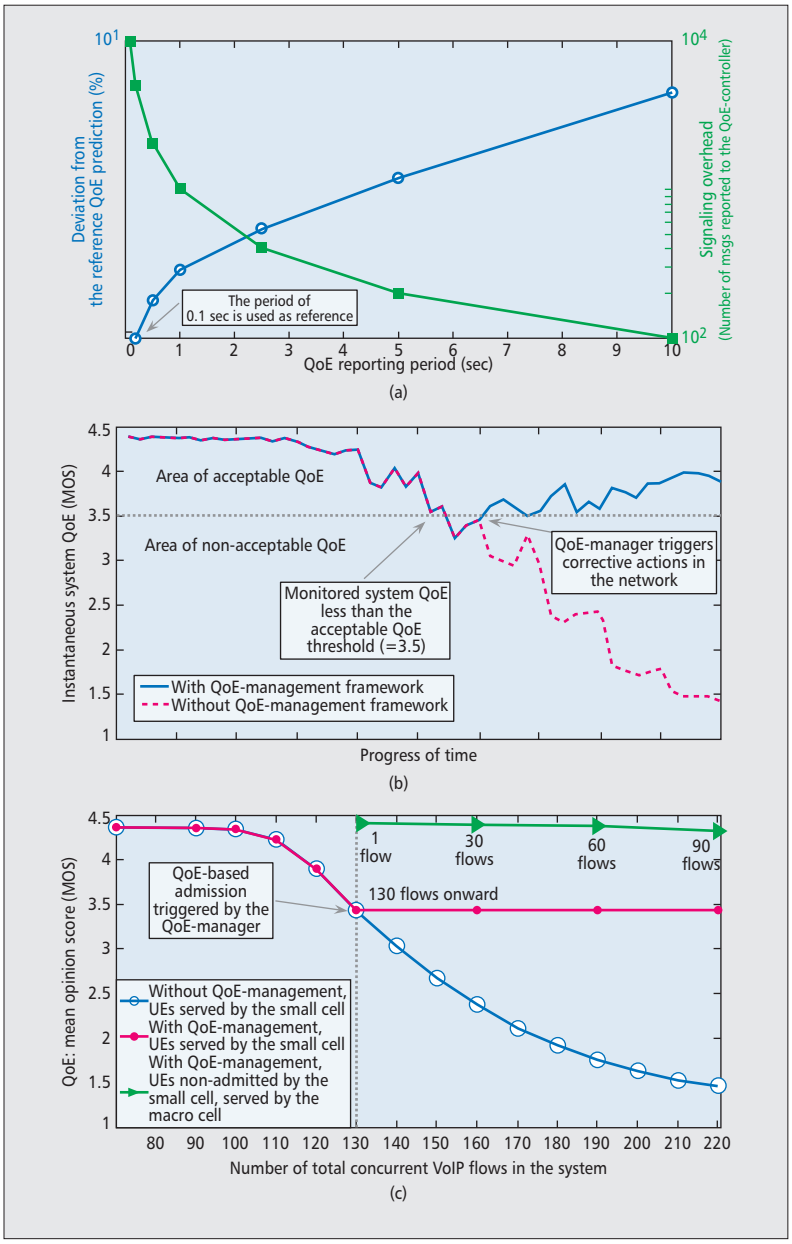


Figure 4. Evaluation results: a) Trade-off between the network overhead and achieved accuracy in the QoE prediction; b) Real-time operation of the QoE management framework; and c) QoE-driven admission control.

As a characteristic example, we describe how the proposed framework may be customized and applied toward implementing a real-time QoE-aware admission controller. We study the case of a heavily congested outdoor small-cell, representing for instance scenarios where this small-cell is used to serve a stadium during a concert or football game. We evaluate the proposed QoE management framework and compare it with the conventional case, where in the absence of QoE awareness, users are admitted based on their positions or on received signal strengths from surrounding base stations. The proposed framework is customized as follows.

QoE-monitor: We study the case of UEs producing VoIP traffic, and select the E-model implementation for the purposes of QoE estimation. Thus, the QoE-monitor provides the QoE-manager with real-time estimations of the QoE experienced per VoIP flow.

QoE-controller: The data collection procedure is tuned, a priori, with the QoE modeling function. Consequently, the E-model dictates the periodic collection of:

- The average delay associated with the transmitted packets, extracted through examining the timing information available inside the received packets.
- The packet loss rate, estimated as the number of erroneously received packets over the aggregate number of transmitted packets, measured by the number of negative acknowledgments produced throughout the QoE reporting period.
- The packet loss robustness factor (the average number of consecutively lost packets over this number for the case of random loss), acquired using statistical information by intermediate network nodes.
- The codec type of the UEs, required to select the appropriate E-model coefficients.

QoE-manager: The QoE-manager is informed by the QoE-monitor about the estimated QoE per VoIP flow, and consequently is aware of the average QoE of the served UEs. If this QoE score reaches a minimum acceptable threshold (here, MOS = 3.5), the QoE-manager will restrict the admission of new flows inside the small-cell. Instead, those will be served by the macro-cell. In this way, a QoE-driven admission control mechanism is implemented.

To evaluate this framework, we generate a constantly increasing number of VoIP flows inside the small-cell, namely within the range of the HeNB, using the simulation parameters of Table 2, and we record the instantaneous average QoE in the system, while time progresses (Fig. 4b). We observe a point when this QoE drops below the predefined MOS threshold, due to the increasing number of competing requests for spectrum resources. This event triggers the QoE-manager to restrict the admission of new flows inside the small-cell, causing any newcomers to be admitted by the macro-cell instead. If the macro-cell is not severely congested, as is the case here, the average system QoE will be lifted above the threshold (blue plot in Fig. 4b), which is not the case if this QoE admission mechanism is not present (red plot).

In Fig. 4c we look at the same experiment in a

more microscopic level, namely we evaluate the achieved QoE level for users admitted either by the small-cell or the macro-cell. Again, we observe at some point a QoE drop below the threshold (specifically, for 130 concurrent VoIP flows inside the small-cell). At this point the QoE-manager does not allow any new flows to be admitted by the HeNB, and so the average QoE inside the small-cell remains constant onward (red plot in Fig. 4c). In parallel, the new flows, which are forced to be served by the eNB, also receive good QoE (green plot), subject, however, to the current load of the macro-cell (note a small QoE decrease from 1 to 90 admitted flows). Consequently, we conclude that the application of this QoE management framework surpasses conventional admission control schemes, which would force all new flows to associate to the HeNB based on QoE-unaware criteria (blue plot).

Conclusions

Mobile cellular technologies, such as 4G and 5G, are moving from network-centric to user-centric approaches, by incorporating some kind of QoE logic and intelligence. Toward this direction, this article focuses on the integration of QoE acquisition and QoE management inside these networks. A framework for end-to-end QoE management is proposed, its viability is investigated, and key challenges for its realization are identified and discussed. Therefore, this work contributes to the need to provide more structured and focused insight on the issue of QoE management in mobile cellular networks, assisting operators with QoE aspirations to adopt this framework and customize it according to specific requirements and needs.

Acknowledgments

This work is supported by the European Commission under the auspices of the FP7-PEOPLE MITN-CROSSFIRE project (grant 317126).

References

[1] Qualinet European Network on Quality of Experience in multimedia systems and services, "Definitions of Quality of Experience (QoE) and related concepts," White Paper, 2012.

[2] R. Stankiewicz, P. Cholda, and A. Jajszczyk, "QoX: What is it Really?" *IEEE Commun. Mag.*, vol. 49, no. 4, Apr. 2011, pp. 148–58.

[3] J. Zhang and N. Ansari, "On Assuring End-to-End QoE in Next Generation Networks: Challenges and a Possible Solution," *IEEE Commun. Mag.*, vol. 49, no. 7, July 2011, pp. 185–91.

[4] G. Gómez *et al.*, "Towards a QoE-Driven Resource Control in LTE and LTE-A Networks," *J. Comput. Networks Commun.*, vol. 2013, Article ID 505910, 2013, pp. 1–15.

[5] A. Cuadra-Sanchez *et al.*, "A Global Customer Experience Management Architecture," *Proc. Future Network & Mobile Summit* (FutureNetw), Berlin, 2012, pp. 1–8.

[6] S. Möller *et al.*, "Speech Quality Estimation: Models and Trends," *IEEE Signal Process. Mag.*, vol. 28, no. 6, Nov. 2011, pp. 18–28.

[7] J. Seppänen, M. Varela, and A. Sgora, "An Autonomous QoE-Driven Network Management Framework," *J. Vis. Commun. Image Represent.*, vol. 25, no. 3, Apr. 2014, pp. 565–77.

[8] S. Baraković and L. Skorin-Kapov, "Survey and Challenges of QoE Management Issues in Wireless Networks," *J. Comput. Networks Commun.*, vol. 2013, article ID 165146, 2013, pp. 1–28.

[9] A. Takahashi, D. Hands, and V. Barriac, "Standardization Activities in the ITU for a QoE Assessment of IPTV," *IEEE Commun. Mag.*, vol. 46, no. 2, Feb. 2008, pp. 78–84.

[10] S. Jelassi *et al.*, "Quality of Experience of VoIP Service: A Survey of Assessment Approaches and Open Issues," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, Jan. 2012, pp. 491–513.

[11] R. Schatz *et al.*, "From Packets to People: Quality of Experience as a New Measurement Challenge," *Data Traffic Monitoring and Analysis*, vol. 7754, E. Biersack, C. Callegari, and M. Matijasevic, Eds. Heidelberg: Springer Berlin, 2013, pp. 219–63.

[12] P. Rengaraju *et al.*, "On QoE-Monitoring and E2E Service Assurance in 4G Wireless Networks," *IEEE Wireless Commun.*, vol. 19, no. 4, Aug. 2012, pp. 89–96.

[13] D. Tsolkas *et al.*, "The Need for QoE-Driven Interference Management in Femtocell-Overlaid Cellular Networks," *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, vol. 131., I. Stojmenovic, Z. Cheng, and S. Guo, Eds.: Springer International Publishing, 2014, pp. 588–601.

[14] P. Reichl, B. Tuffin, and R. Schatz, "Logarithmic Laws in Service Quality Perception: Where Microeconomics Meets Psychophysics and Quality of Experience," *Telecommun. Syst.*, vol. 52, no. 2, Feb. 2013, pp. 587–600.

[15] G. Piro *et al.*, "Simulating LTE Cellular Systems: An Open Source Framework," *IEEE Trans. Vehic. Tech.*, vol. 60, no. 2, Feb. 2011, pp. 498–513.

Biographies

Eirini Liotou (eliotou@di.uoa.gr) received a Diploma in electrical and computer engineering from the National Technical University of Athens, and the MSc in communications and signal processing from the Imperial College of London. She has worked as a senior software engineer at Siemens Enterprise Communications within the R&D department. Since 2013 she has been a researcher and Marie Curie fellow in the Department of Informatics & Telecommunications at the University of Athens, working on QoE provisioning in 4G/5G networks.

Dimitris Tsolkas (dtsolkas@di.uoa.gr) received the B.S. degree in informatics and telecommunications and the M.S. degree in communications systems and networks from the Department of Informatics and Telecommunications, University of Athens, Greece, in 2007 and 2009, respectively. In 2014 he received his Ph.D. degree from the same department. Currently he works as a researcher in the Green, Adaptive and Intelligent Networking Group within the Dept. of Informatics and Telecommunications. His research interests include radio resource management, D2D communications, and QoE provisioning.

Nikos Passas (passas@di.uoa.gr) received his diploma (honors) from the Department of Computer Engineering, University of Patras, Greece, and his Ph.D. degree from the Department of Informatics and Telecommunications, University of Athens, Greece, in 1992 and 1997, respectively. Since 1995 he has been with the Communication Networks Laboratory of the University of Athens, working as a senior researcher in a number of national and European research projects. His research interests are in the area of mobile network architectures and protocols.

Lazaros Merakos (merakos@di.uoa.gr) received the diploma in electrical and mechanical engineering from the National Technical University of Athens, Greece, in 1978, and the M.S. and Ph.D. degrees in electrical engineering from the State University of New York, Buffalo, in 1981 and 1984, respectively. From 1983 to 1986 he was on the faculty of the Electrical Engineering and Computer Science Department, University of Connecticut, Storrs. From 1986 to 1994 he was on the faculty of the Electrical and Computer Engineering Department, Northeastern University, Boston, MA. During the period 1993–1994 he served as Director of the Communications and Digital Processing Research Center, Northeastern University. During the summers of 1990 and 1991 he was a visiting scientist at the IBM T. J. Watson Research Center, Yorktown Heights, NY. In 1994 he joined the faculty of the University of Athens, Athens, Greece, where he is presently a professor in the Department of Informatics and Telecommunications, and Scientific Director of the Networks Operations and Management Center.

This work contributes to the need to provide more structured and focused insight on the issue of QoE management in mobile cellular networks, assisting operators with QoE aspirations to adopt this framework and customize it according to specific requirements and needs.

An Open Framework for Programmable, Self-Managed Radio Access Networks

Kostas Tsagkaris, George Poulios, Panagiotis Demestichas, Abdoulaye Tall, Zwi Altman, and Christian Destré

ABSTRACT

SON technologies have been a promising approach for simplified and automated RAN management. Currently, SON functions can be D-SON among the network elements or C-SON as part of the OSS/NMS. D-SON technology is typically provided by equipment vendors, while C-SON is provided by both third parties and equipment vendors. D-SON has high reactivity but is usually provided as black boxes with poor or no interoperability between different vendors and management systems, hence rendering their integration a challenging task. On the other hand, OSS/NMS-level SONs, although provided with a global network view, are limited to coarse time operation. In this work we apply SDN principles to the management of RANs and propose a software-defined SON (SD-SON) framework that tackles the shortcomings and combines the benefits of both approaches above, namely, high reactivity along with flexibility, programmability, and openness. The architectural components, implementation aspects, and advantages of the proposed SD-SON framework are described in detail. A proof-of-concept prototype of a programmable self-managed LTE-Advanced heterogeneous network is also presented.

INTRODUCTION

A self-organizing network (SON) is a radio access network (RAN) equipped with self-* management capabilities [1]. After being introduced into the first release (8) of LTE, SON technology continued to evolve and to enrich LTE and LTE-Advanced technologies. Mobile operators have shown a particular interest in SON for different reasons. First, RANs are becoming increasingly complex and heterogeneous with co-existing and co-operating technologies. Second, operators evolve in an increasingly competitive eco-system and face continuous erosion in operational margins. Furthermore, the permanent increase in traffic demand forces operators to continuously invest in network infrastructure. In this context, operators need to reduce operational expenditure.

SON technology alleviates operational costs and improves network performance and profitability. It makes it possible to autonomously configure newly deployed network nodes (self-configuration), tune parameters to improve key performance indicators (KPIs[1]) (self-optimization), and perform diagnostic and reparation of faulty network nodes (self-healing). Two types of solutions in the process of field trials and first large scale deployments are being developed today. One solution is distributed SON (D-SON), with SON functions installed within eNodeBs. This solution can be seen as *control plane SON* and is proposed by infrastructure vendors. It has the advantage of supporting reactive operation of SON functions, for example, with time scale of the order of a minute. The other solution is centralized SON (C-SON), with SON functions deployed in a server on top of the element management system (EMS), namely in the network management system (NMS)/OSS of the operator. C-SON can be seen as a management plane solution and is developed by both third parties and by infrastructure vendors. C-SON can benefit from abundant data at the EMS, but has limited reactivity, with a minimum periodicity typically set to half or one hour. Figure 1 depicts the D-SON and C-SON solutions.

In this work we claim that a solution that addresses the shortcomings and combines the benefits of both C-SON and D-SON can be found by applying Software Defined Networking (SDN) principles [2] to the management of RANs. Accordingly, we propose a reference Software-Defined SON (SD-SON) framework for control plane SON. SD-SON will benefit from the control plane high reactivity of D-SON, that is, short time scale variations in traffic and propagation conditions, while as in C-SON it will yield benefits from a) short implementation cycles, in contrast to software upgrades in RAN, typically occurring once or twice a year; and b) flexibility to introduce and evolve new SON and management functions independently of standardization.

In accordance with SDN principles, the framework offers a logically centralized entity called the SD-SON controller, which

- Maintains a global view of the underlying network.
- Provides well-defined application programming interfaces (APIs) for developing and introducing SON functions by anybody and for dynamically programming their activation/deactivation and configuration on the fly.

Kostas Tsagkaris, George Poulios, and Panagiotis Demestichas are with University of Piraeus.

Abdoulaye Tall, Zwi Altman, and Christian Destré are with Orange Labs.

[1] "KPIs" is a standard term used in RAN and SON management, that provides network operators with a comprehensible overview of the otherwise vast and technically complex performance metrics namely, Quality of Service or Experience (QoS and QoE), OPerational EXpenditures (OPEX), revenue maximization and more.

0163-6804/15/$25.00 © 2015 IEEE

• Allows for SON function separation and isolation for multiple third party SON providers.

Moreover, in order to achieve these goals an appropriate abstraction similar to that of the "flow" in typical data-center SDNs is required. The key choice of abstraction in SD-SON are the SON *function*, and its input and output (I/O), namely the *configuration parameter*, and the *metric* (KPIs).

SD-SON forms a common control substrate and enables cross-vendor SON programmability. SON functions can be provided by infrastructure vendors, third parties, or the operator R&D, as software upgrades in a vendor agnostic manner, thus offering an opportunity to escape strong vendor dependencies and to tackle EMS heterogeneity. This open and programmable environment presupposes that vendors are also willing to "open" their equipment and comply with the standards of the proposed SD-SON framework. We believe this is a reasonable expectation for future RAN systems, considering their ever growing complexity and cost, and the current trend toward softwarized networks and openness. Cloud-RAN [3] is just one example where the SD-SON framework could naturally fit.

Related Work

SDN and flow-networking is not as significant in RANs as it is in data-centers and core networks that are mainly formed by switches and routers. However, OpenRoads [4] adopts the OpenFlow protocol as well as FlowVisor for the slicing of mobile networks. Its main purpose is to enable concurrent experimentation and to tackle mobility issues such as the seamless handoff over different radio access technologies. For the configuration of network elements, OpenRoads employs SNMP and a custom virtualization layer, which in combination with FlowVisor allows experimenters to have virtually full control of the network.

On a different track, the SoftRAN architecture [5] proposes a logically centralized control plane through the abstraction of the so called "Big Base Station" encapsulating all base stations in a RAN. Similar to CloudRAN [1] and Cloudiq [6], it leverages the global network view to simplify the implementation of the otherwise non-trivial, network-wide functionalities, such as load balancing, utility maximization, and Layer 1 coordination. SoftCell [7] targets the simplification of cellular networks by, among other methods, placing middleboxes in the cellular core and supporting high-level, subscriber-oriented policies that dictate the traversal of traffic through those middleboxes. OpenRadio [8], on the other hand, provides a rich, wireless protocol programming interface of rule-action and state machine models, similar to software radio.

Such approaches attempt to introduce SDN to wireless networks. However, so far no focus has been placed on SON programmability. The contribution of SD-SON lies in the definition and incorporation of new RAN-oriented and SON-oriented abstractions in both the control plane and the management plane.

Furthermore, management of SON can be seen as a special case of autonomic network management (ANM). ANM applies the 'monitor, analyze, plan, execute' model of IBM's autonomic computing [9] in network management and has been extensively adopted and used by several research endeavours in the past. Only a few focused particularly on SON [10], whereas the majority focused on the definition and standardization of generic frameworks and common adaptation layers for managing self-* functions in an abstract way regardless of their context of application. The Unified Management Framework (UMF) [11] and the holistic Generic Autonomic Network Architecture (GANA) standardized by ETSI AFI [12] were among the most recent ones.

Our work is greatly influenced by these architectures, and especially by the UMF [11], which we adopt and enhance by applying SDN principles and by narrowing its scope to the management of SON functions in RAN, toward our SD-SON framework. Thus, although the concept of a common adaptation layer is not novel, its implementation for vendor-agnostic SON operation is very recent and it applies to the first significant large scale instantiation of autonomic management, i.e. SON.

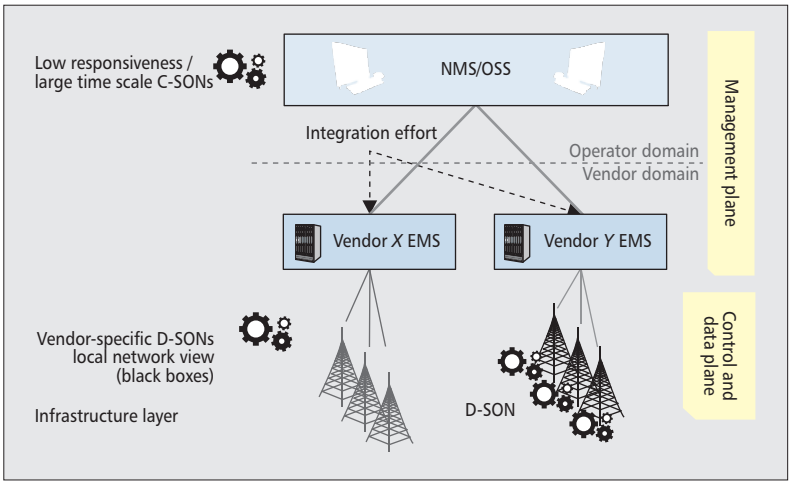


Figure 1. Typical positioning of SON functions in the control and management planes.

SD-SON Framework

High-Level Architecture

The main functional blocks of the SD-SON architecture are the *management core*, the *SD-SON controller*, and the *SON functions* (hereafter, *SONs*). Figure 2 depicts an overview of these components and their interfacing. From the bottom up, network elements are required to comply with the southbound (SB) interface of the controller, i.e. register their parameters and metrics and update them accordingly. The controller is then responsible for exposing several actions on those elements through its northbound (NB) interface. Such actions include, but are not limited to, element discovery, generation of events regarding metric values, and reading or writing of parameters. SONs may use the NB interface to apply their (potentially network-wide) functionality in a vendor-agnostic way. For their management, SONs are required to implement a management interface that is common

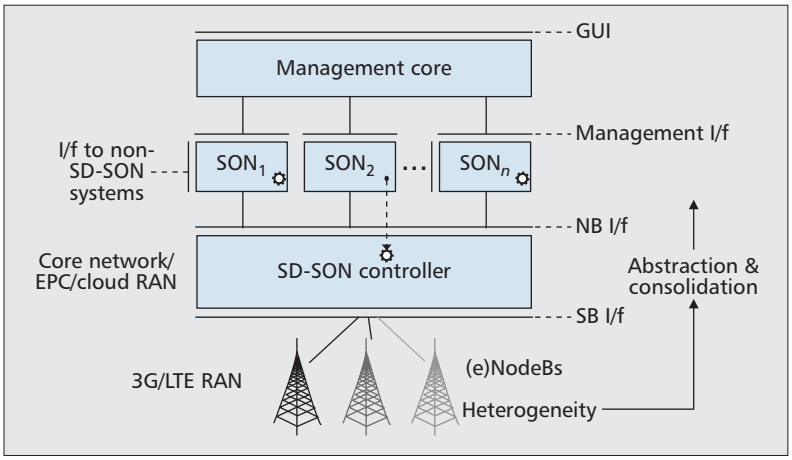


Figure 2. High level SD-SON architecture.

regardless of their nature. Among others, this interface is used to start or stop the SONs, apply policies, and deploy them over particular network elements.

In terms of practical deployment, SD-SON can be deployed in a new network element or in an existing one such as the mobility management entity (MME) within the core network. A particularly attractive implementation of SD-SON would be within the cloud-RAN. SD-SON can be useful for all types of SON mechanisms: self-configuration (e.g. switching eNodeBs on and off for energy saving), self-optimization (e.g. interference mitigation, load balancing and off-loading, coverage capacity optimization), or self-healing (e.g. cell outage compensation).

An important aspect of SONs is that their functionality (little gear in Fig. 2) is not necessarily embodied by them, as in SON_1. Alternatively, as indicated by SON_2, the main code of the SON function may be downloaded into the controller along with the required accompanying meta-data such as the inputs, output, and the criteria for invocation. Given this information, the controller then undertakes the linkage of the function with its I/O and its execution according to the specified invocation criteria (e.g. every 60 seconds or whenever a metric changes). The choice of whether to embody or download the function logic into the controller is left up to the SON designer and depends on several factors such as the interaction with non-SD-SON systems (westbound), computational power requirements, time-scale constraints, and more. In particular, downloaded functions (as in SON_2) enjoy higher reactivity, and fewer remote method invocations (RMI), while they may have limited interaction with unsupported SD-SON interfaces/libraries. The opposite holds for the standalone SONs (as in SON_1), which enjoy complete programming and interfacing freedom at the cost of REST-RMI implementation and delay overhead.

Information Model

Metrics, Parameters, and Elements: No assumption is made in the SD-SON information model about the types of parameters, metrics, or network elements. The design pattern followed considers a *specification* (henceforth used interchangeably with *type*) for each *instance* of a concept and is heavily influenced by TeleManagement Forum's shared information/data (SID) framework [13], hence making the framework easy to integrate with legacy management systems (OSS/NMS). Figure 3 is a simplified class diagram of these concepts.

Metrics and *parameters* are similar in that they both consist of a specification, an actual value handler (i.e. the metric source and parameter adapter, respectively), and an associated con-

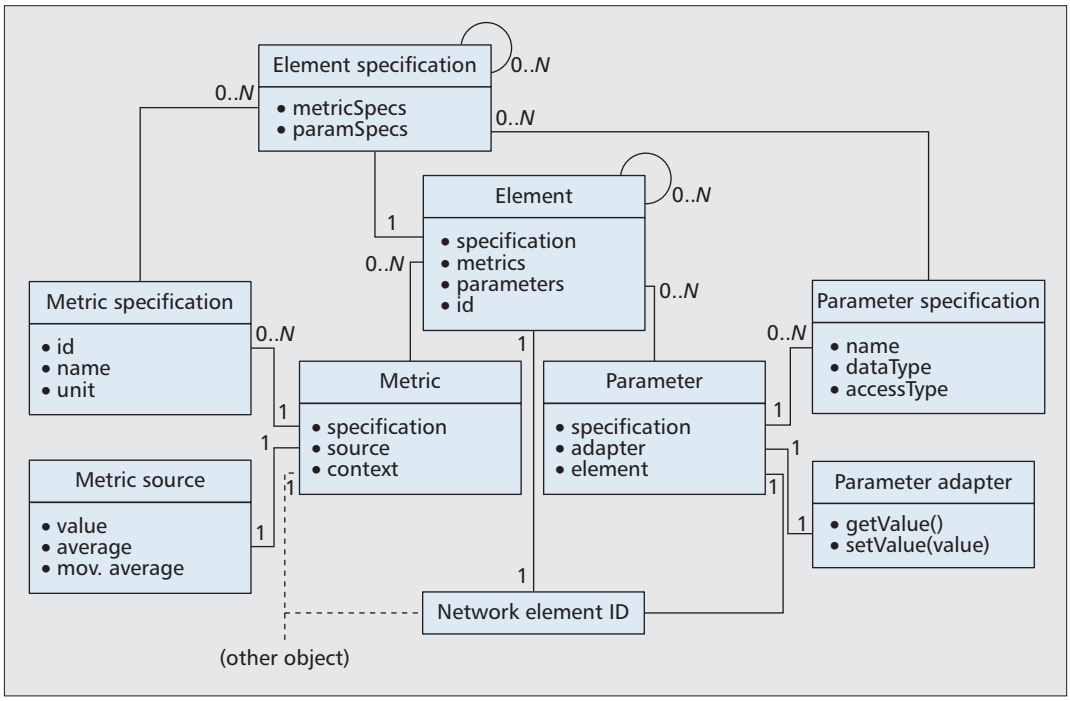


Figure 3. Metric, parameter, and element model of SD-SON framework.

text/element. Many metrics or parameters may have the same specification (1-to-N relation) while each of them must be associated to a different context/element (1-to-1 relation). For metrics, the context may be pointing to a network element or some other entity such as a subnet, or a cell cluster. On the other hand, parameters are always associated to a network element (physical or virtual).

The *metric source* is the endpoint for acquiring metric values. It is not strictly required to be the producer of the data, but rather the logical source in the scope of the SD-SON framework. The rationale behind it is to expose a common data-retrieval interface regardless of the provider (protocol/API) or the type of metric. Metric sources also generate various kinds of events regarding the state of their value with respect to a range, a time interval, or the previous value. Such events may be used, among others, to trigger the invocation of SON functions or selectively retrieve metric data.

Similar to the metric source, the *parameter adapter* provides a common interface to read or write a parameter value. Adapters may be implemented for any kind of configuration protocol/API (e.g. SNMP or NETCONF), while depending on the access type of their corresponding parameter specification, they may allow read-only access, read-write by a single entity at a time, or read-write by multiple entities. This distinction between write access types is also used by the controller to prevent write conflicts on parameters as deemed necessary by their specification. More sophisticated mechanisms can also be introduced to the controller to resolve conflicting or inconsistent parameter settings, depending on the deployment scenario per se.

Elements, in turn, are a conglomeration of metrics, parameters, and other elements. This recursive definition (elements within elements) allows for reusability of types of equipment at finer granularities. For instance, the antenna or remote radio head (RRH) of a base station (BS) can be considered as elements of the greater BS-element. In this way, BSs with potentially different specifications may share the same type of antenna or RRH with the same parameters.

SON Function: The three most important parts of a function definition are its inputs, output, and body/code. Besides the specifications of metrics or parameters taken as inputs and given as output, it is required to specify which particular instances of metrics and parameters (equivalently, which elements) to apply the function to, along with when and how often to invoke/apply it. For the former, SD-SON defines a list of tuples called *invocations* accompanying any function. Each of them specifies the elements or context of the parameters or metrics respectively, to use as inputs and output on a single invocation. For the latter, functions are also accompanied by a set of *triggers* specifying when should a function be applied. Triggers may be based on either timers, metric events, or both.

All the aforementioned properties are encapsulated by a data structure called SONFunction and can be passed to the controller through the NB interface for installation and execution during runtime (as in Fig. 2, SON_2). Alternatively, they may be embodied within an external hosting entity (as in Fig. 2, SON_1), in which case the NB would be used only for getting/setting the required/computed values.

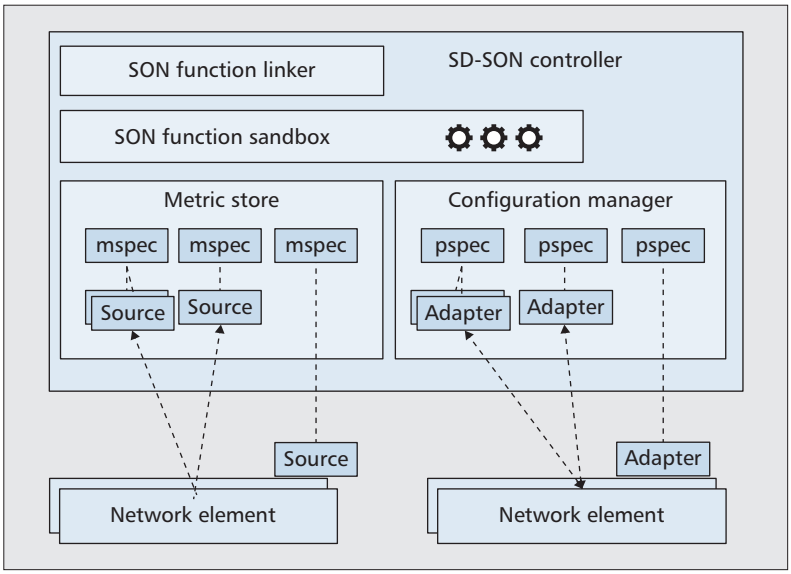


Figure 4. Functional overview of the SD-SON controller.

SD-SON Controller

Through the SD-SON controller, and in particular its NB API, SONs can:

- Discover existing network elements and topology.
- Retrieve or configure the value of parameters.
- Monitor metrics or subscribe to events related to metric value changes.

On the other side, network elements, metrics, parameters, and their associations can be introduced into the system through the SB interface of the controller. Metric and parameter values can then be updated through the SB while the controller undertakes the synchronization with any NB parameter writings.

Figure 4 depicts a functional overview of the controller. The *metric store* is responsible for maintaining an inventory of metric specifications and sources. It is also the endpoint for registering such specifications ("mspecs" in Fig. 4) and the associated sources. The latter may reside either inside the controller or the infrastructure layer, enabling both centralized and distributed data dissemination schemes, since the SONs can retrieve data directly from them. In both cases, each source is registered with one of the specifications in the controller so that interested entities may discover it. In order to minimize the network footprint, bulk data may be pushed into multiple sources at a time, in the case these are maintained inside the controller.

The *configuration manager* keeps track of the parameters, provides an interface to their adapters, and ensures their access restrictions are respected by the NB entities, according to the parameter specifications ("pspecs" in Fig. 4).

The *SON function linker* undertakes compilation and linkage of SONFunction objects to their

```
<son-function name="CIO Load Balancing">
  <inputs>
    <metric id="load" value="CURRENT"/>
    <metric id="load" value="CURRENT"/>
  </inputs>
  <output>CIO</output>
  <invocations>
    <invocation>
      <inputs>pico3, macro0</inputs>
      <output>pico3</output>
    </invocation>
    <invocation>
      <inputs>pico4, macro0</inputs>
      <output>pico4</output>
    </invocation>
  </invocations>
<code> <![CDATA[
  function(cio, picoLoad, macroLoad) (
    var e = 0.05;
    var delta = e * (macroLoad-picoLoad) ;
    return Math.min(Math.max(cio+delta, 0), 12.0) ;
  }
]]></code>
  <triggers>
    <trigger type="METRIC_CHANGE_DRIVEN" metricid="load"/>
  </triggers>
</son-function>
```

Figure 5. An XML description of the load balancing SON function.

specified inputs, outputs, and triggers. Linked functions are then forwarded to the *SON function sandbox* for execution. Sandboxing is necessary because functions are presented at runtime, and executed within the controller's process space. Hence, the sandbox is in practice a scripting engine, restricting I/O access to other functions, the controller internals, or the host operating system, thus preserving controller stability in case of either intentional or accidental misbehaving or uninterpretable (e.g. syntactically wrong) code of a SON. Javascript was chosen for coding SON functions, merely because of its smooth integration/interoperability with Java (in which the controller itself runs). Any other language could have been used instead with a minimum requirement to be able to load code at runtime.

Communication Model

Interfacing in the reference implementation of the SD-SON framework is achieved through a custom Java RMI mechanism which, in contrast to the traditional one, uses HTTP to perform remote method invocations. Using this approach, integration time of different components is minimized while both remote SONs and network elements can be introduced to the system at runtime. From the developer's point-of-view, using RESTful APIs, whether as a server or a client, happens completely seamlessly in the case of using Java, i.e. because mapping of methods and de/serialization (a.k.a. un/marshalling) of objects between Java and HTTP are transparently undertaken by the underlying RMI mechanism.

The data formats used vary from plain text to XML and JSON. Although the framework defaults to one of them when performing Java RMIs, developers are free to post and request any particular data format at will, as long as they respect the specified data structure.

Instantiation to 4G LTE-A

Storyline

According to 4G LTE-A, small cells (e.g. pico cells) can be used to offload macro cell traffic to increase RAN capacity. This is often denoted as load balancing (LB) among the macro cell and the small cells. The latter can be deployed close to the macro cell edge, where typically more macro cell resources are needed to serve UEs, and/or at hot spots.

To enhance off-loading capability, 3GPP has introduced two jointly operating self-optimization mechanisms [14]. The first is called cell range extension (CRE) and aims at extending the coverage of small cells by increasing their cell individual offset (CIO) parameter. The CRE is optimized using a LB SON. However, the traffic at the extended coverage zone will suffer from interference coming from the macro cell. Therefore, a second mechanism is introduced, namely, an interference mitigation which, on certain time intervals of 1 ms, a.k.a. subframes, mutes almost all macro BS transmissions. These subframes are denoted as almost blank subframes (ABS), and the ABS SON adjusts the ratio of muted-to-transmission subframes for the macro BSs. Further technical details of the functioning of the two SONs can be found in [14, 15].

In order to illustrate the presented SD-SON concepts, both SON mechanisms have been developed as a generic implementation, while ignoring the particularities of a vendor implementation for the eNodeB, EMS, NMS/OSS, or any of their interfacing. The only assumption made is that, given a SON development kit (including the required SD-SON libraries and APIs), a SON provider develops a function and distributes/sells it as a hot-pluggable software package (i.e. an SDN application). The instantiated framework is responsible for providing smooth adaptation for such SONs (through the SDN-SON controller), and driving them through their life-cycle (through the management core) under the ultimate command of the network operator empowered with real-time monitoring insights.

Network Registration and SON Function Programming

The first step is to introduce the network elements to the SD-SON controller through its SB API. Two element specifications will suffice in this 4G LTE-A use case: one for the macro BS and one for the pico BS, although a finer decomposition into sub-elements (e.g. into antennas and so on) could also be considered. In order to provide an element specification, one needs its parameter and metric specifications (pspecs and mspecs in Fig. 4, respectively). In turn, element specifications for macro and pico BS may be registered as a conglomeration of several parameters and metrics, for example, with parameter types for transmission/pilot power, location, ABS ratio, and CIO, and metric types for load, throughput, blocking rate, and so on. The next

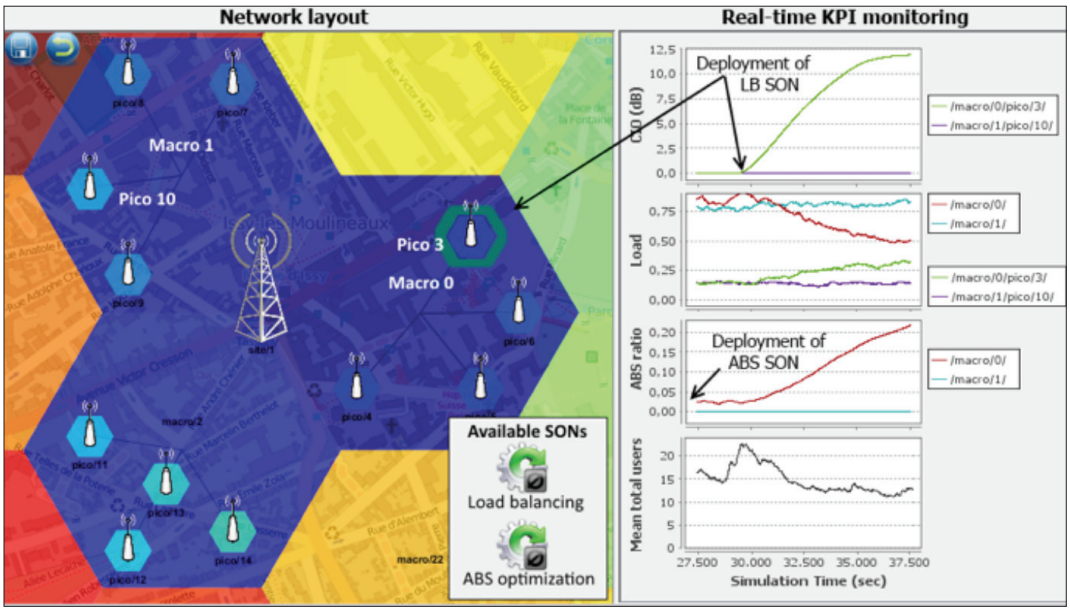


Figure 6. Screenshot of the SD-SON prototype during the deployment of SONs

In order to illustrate the presented SD-SON concepts, both SON mechanisms have been developed as a generic implementation, while ignoring the particularities of a vendor implementation for the eNodeB, EMS, NMS/OSS, or any of their interfacing.

step would be to implement metric sources and parameter adapters and register them to the controller. Through the SB, network element instances can then be created and associated with their previously registered specification, their metrics and parameters.

Following the registration of the network elements to the controller, SONs may use the NB API to retrieve the required KPIs and configure their targeted parameters. Figure 5 is a simplified, XML-formatted description of the CIO-based LB SON presented earlier. Such a description would be installed through the NB API for execution inside the controller, unless embodied within the external SON process.

The first XML node specifies the type of inputs and output of the function. The LB SON takes as input the "CURRENT" value of two "load" metrics and configures as output the parameter with name "CIO." The next node, "invocations," is a list with tuples specifying the particular elements (see child nodes) whose loads is to be read from, as well as the elements whose CIOs should be configured. In this example, the function will be invoked twice (i.e. for pico3 and pico4) per triggering.

The following "code" node contains the logic of the SON to be executed per iteration. This field may only contain a single Javascript function with arbitrarily named arguments. The following three lines of code correspond to equation (2) in [15] with an additional minimum and maximum value limitation. The measurement unit of CIO is dB. The last node named "triggers" contains the criteria for application of the function. It specifies that it should be invoked whenever the load changes. The ABS SON would be similar in structure but applied on a per-macro BS basis.

Proof-of-Concept Prototype

An SD-SON proof-of-concept prototype has been implemented and applied on a 4G LTE-A simulated network according to the aforementioned use case. Twenty-one macro and 12 pico BSs were simulated, out of which 18 macro BSs were used to simulate surrounding cell interference only. Figure 6 is a screenshot of the management core dashboard during the deployment of the ABS and LB SONs. On the left hand side, the blue sectors depict the three macro and their 12 pico cells considered for the SON deployment, namely macro/0, macro/1, and macro/2, and pico/3 through pico/15.

For comparison purposes, the SONs have been deployed only on macro/0 and its picos/3-6. On the right hand side of Figure 6, a real-time KPI and parameter evolution monitoring panel illustrates the effects of the SONs. The red curves correspond to macro/0 and the turquoise curves to macro/1. Green curves correspond to pico/3 (residing in macro/0 cell); blue curves correspond to pico/10 (residing in macro/1 cell). The horizontal axis is the simulation time in seconds; the verticals are, from top-down, the evolution of CIO, load, ABS ratio, and mean number of users in the system.

The initial load is approximately 80 percent for the macros and 20 percent for the picos. The ABS SON is deployed over macro/0 early during the simulation and maintains low values of approximately 0.025 since the macro BSs serve most of the traffic and therefore the ABS SON keeps the muting ratio of the macro subframes low. Due to the imbalance in the loads of macros and picos, by the time the LB SON is deployed over picos of the macro/0 cell, it starts increasing their CIO. Indicatively, only the CIO of pico/3 has been plotted, as shown in the top chart of Fig. 6. The next chart shows how a portion of the load of macro/0 is off-loaded to its respective pico cells (once again only pico/3 is plotted for clarity), in contrast to the macro/1 cell, which is operating without SONs. The third chart depicts how the ABS SON increases the ratio of blank subframes as the traffic served by the pico cells increases due to the off-loading. Last but not least, the bottom chart indicates a lower average

By using bulk data transfers on the SB and well-packed JSON objects (without compression), and taking into account the periodicity of the SON function of the order of a minute, and the low amount of data required for its operation, only a small portion of the time was spent interacting with the controller.

number of users in the system due to the better distribution of load, which results in a lower service time of elastic traffic.

Thanks to the SD-SON framework, SONs are offered as distributable software packages, deployed on-the-fly and without any assumptions made about the BS/EMS vendor or interfaces. That is, the same exact SON technology could be deployed over a different real-life network, only by porting the required parameter adapters and metric sources.

In terms of efficiency and scalability, the simulation was running in around 120 times faster than real-time. The controller was run on a 2.4 Ghz CPU core and communicating over WLAN with the simulator, and yet, the bottleneck was the computing power of the machine running the simulation. By using bulk data transfers on the SB and well-packed JSON objects (without compression), and taking into account the periodicity of the SON function of the order of a minute, and the low amount of data required for its operation, only a small portion of the time was spent interacting with the controller. This observation indicates that in a real-life pace, the proposed system can be efficient enough.

SD-SON in Cloud RANs

Cloud-RAN [1] capitalizes on advances in optical and virtualization technologies to enable massively centralized BS deployment by enabling the connection of thousands of RRHs to a centralized, cloud-hosted, pool of baseband units (BBU). Besides the benefits of hosting BBUs in the cloud (*pay-as-you-go*, easier deployment and updates, capability to support multiple standards), two additional advantages are identified from a potential interplay between SD-SON and Cloud-RAN:

(a) BBUs are co-located, thus they are physically capable of collaborating.

(b) BBUs are considered to be hosted on commodity, a.k.a. commercial, off-the-shelf (COTS), hardware (typically x86/ARM-based servers) and not provided from the vendor of the BS.

For SD-SON, (a) basically solves the issue of placement and distribution of the controller while it is a major boost to scalability. That is, one instance of the controller may logically reside in the same cloud host as the one for the BBU pool. Their interfacing then would be greatly simplified since the BBU pool would be able to provide access to thousands of elements from a single point, hence facilitating the centralized nature of the SD-SON controller. Of course, although this physical/logical view seems straightforward, such a logical interfacing is still to be defined. On the other hand, (b) is obviously aligned with one of the major SD-SON goals/benefits: *vendor-neutrality* and *openness*. Given such a vendor-agnostic BBU pool, the SD-SON controller SB can be greatly simplified from the common adaptation layer provided.

Overall, for the reasons outlined above, integrating the SD-SON SB with a cloud-RAN-based architecture, whether fully centralized (layer 1-3 functionality in BBU) or partially (layer 2-3 functions in BBU), seems highly beneficial as both approaches are aligned when it comes to centralization and vendor independence. Further to that, transport issues at the SD-SON SB (such as latency and network overhead) during the transmission of metrics and the configuration of parameters are inherently tackled by the cloud-RAN due to the orders of magnitude higher transmission requirements it has (hence the use of optical links between the BBU and the RRHs).

Discussion and Conclusions

A novel control plane framework for programmable and self-managed RANs has been presented in this article. The following substantial benefits resulting from the adoption of the SDN principles in the SON technology are observed:

- Programmable SON functions with highly reactive operation are made possible.
- Control and management of SON functions shift toward the operator domain and challenges vendors to open their equipment.
- Outsourcing of SON functions is enabled, increasing competition and innovation.
- A broader network view for SON functions.
- Extensibility to new KPIs, parameters, and elements (even during runtime) is enabled by avoiding hard-coded models.
- Continuous evolution of the the SON ecosystem (i.e. control and management intelligence).

However, the proposed transitions are not achieved without cost. A challenge is posed to vendors to support and comply with a common interface (i.e. the SB) and to open their equipment. Another challenge is the integration of the SD-SON management core with existing NMS/OSS and BSS. Nevertheless, we claim that both are worth the price and effort and are outweighed by the benefits. Furthermore, future needs of SON deployment and management will likely call for such flexible schemes and architectures. Overall, the SD-SON findings seem capable of transforming the RAN into a smarter and easier to manage system, by significantly facilitating the forecasted massive deployment of the SON technology.

Acknowledgment

This work has been performed in the context of the UNIVERSELF project (Universal Self-management (http://www.univerself-project.eu)), which has been supported by the European Community's Seventh Framework Program (FP7) and the AutoSDN project involving the University of Piraeus Research Center and Orange, under contract no. D02191. This article reflects only the authors' views and the community is not liable for any use that may be made of the information contained therein.

References

[1] L. Jorguseski, "Self-Organizing Networks in 3GPP: Standardization and Future Trends," *IEEE Commun. Mag.*, vol. 52, no. 12, Dec. 2014, pp. 28–34.

[2] K. Hyojoon and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp.114–19.

[3] "C-RAN: The Road Towards Green RAN," white paper from China Mobile, ver 2.5 (Oct. 2011).

[4] K.-K. Yap *et al.*, "OpenRoads: Empowering Research in Mobile Networks," *ACM SIGCOMM Computer Communication Review*, 40.1, 2010, 125–26.
[5] A. Gudipati *et al.*, "SoftRAN: Software Defined Radio Access Network," *Proc. Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, ACM, 2013.
[6] S. Bhaumik *et al.*, "Cloudiq: A Framework for Processing Base Stations in a Data Center." *Proc. 18th Annual Int'l Conf. Mobile Computing and Networking*, ACM, 2012.
[7] X. Jin *et al.*, "SoftCell: Taking Control of Cellular Core Networks," arXiv preprint arXiv:1305.3568 (2013).
[8] M. Bansal *et al.*, "Openradio: A Programmable Wireless Dataplane," *Proc. First Workshop on Hot Topics in Software Defined Networks*, ACM, 2012.
[9] J. O. Kephart and D. M. Chess, "The Vision of Autonomic Computing," *Computer*, 36.1, 2003, pp. 41–50.
[10] R. Litjens *et al.*, "Self-Management for Unified Heterogeneous Radio Access Networks," *Proc. IEEE 77th Vehicular Technology Conf.* (VTC Spring), 2–5 Jun. 2013.
[11] K. Tsagkaris *et al.*, "A Survey of Autonomic Networking Architectures: Towards a Unified Management Framework," *Int'l J. Network Management*, vol. 23, Iss. 6, Nov/Dec 2013, pp. 402–23.
[12] M. Wódczak *et al.*, "Standardising a Reference Model and Autonomic Network Architectures for the Self-Managing Future Internet," *IEEE Network*, vol. 25, no. 6, Nov. 2011, pp. 50–56.
[13] Information Framework (SID) reference page, http://www.tmforum.org/InformationFramework/1684/home.html, TM Forum, July 2013
[14] "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2," 3GPP, TS 36.300 v10.11.0, Sep. 2013
[15] G. Poulios *et al.*, "Autonomics and SDN for Self-Organizing Networks," *Proc. 11th IEEE Int'l. Symposium Wireless Communication Systems* (ISWCS), 2014.

Biographies

Kostas Tsagkaris [SM'] (ktsagk@unipi.gr) holds a Ph.D. (Ericsson awarded) in telecommunications from the School of Electrical and Computer Engineering of NTUA, Greece. Since 2005 he has been a senior research engineer and adjunct lecturer at the University of Piraeus. His research interests include the design, management and optimization of autonomic/cognitive networks, SON and SDN. He has published more than 150 papers in international journals and refereed conferences and also co-founded a startup company in these areas.

George Poulios (gpoulios@unipi.gr) received his undergraduate degree from the University of Piraeus, Department of Digital Systems in 2011, and his masters' degree from the same university in the area of information systems security in 2015. Since 2011 he has been a member of the Telecommunication Networks and Integrated Services (TNS) Laboratory as a research engineer, working on unified, knowledge-based network management.

Panagiotis Demestichas (pdemest@unipi.gr) is a professor at the University of Piraeus, Greece. He is the head of the Digital Systems Department and the director of the M.Sc. degree on technoeconomic management and security of digital systems. He also leads the Laboratory of Telecommunication Networks and Integrated Services. His research interests include wireless/mobile broadband (5G, heterogeneous networks, and cloud RANs), high-speed networking, smart FI infrastructures and services, SDN, and NFV.

Zwi Altman (zwi.altman@orange.com) is a senior research expert at Orange Labs. He received the B.Sc. and M.Sc. degrees at the Technion-Israel Institute of Technology, in 1986 and 1989, respectively, and the Ph.D. degree from the INPT France in 1994. From 1994 to 1996 he was a post-doctoral research fellow at the University of Illinois at Urbana Champaign. He joined Orange Labs in 1996, where he has been involved in projects on network optimization, SON, autonomics, and SDN.

Abdoulaye Tall (abdoulaye.tall@orange.com) received his engineering degree from Tunisia Polytechnic School, La Marsa, Tunisia in 2012. He is currently pursuing a Ph.D. degree in computer science with Avignon University (Avignon, France) under the direction of Zwi Altman (Orange Labs) and Eitan Altman (INRIA). His current research interests include self-organizing networks, queuing theory, convex optimization, and stochastic approximation.

Christian Destre (christian.destre@orange.com) received a Ph.D. degree in computer science from the Evry University (France) in 2004 in the area of operational research applied to all optical network routing. He joined Orange Labs in 2005, specializing in the management of networks and services. Areas of expertise include: autonomic network management, NFV, SDN, and real time OSS evolution.

future needs of SON deployment and management will likely call for such flexible schemes and architectures. Overall, the SD-SON findings seem capable of transforming the RAN into a smarter and easier to manage system, by significantly facilitating the forecasted massive deployment of the SON technology.

On-Demand Scheduling: Achieving QoS Differentiation for D2D Communications

Min Sheng, Hongguang Sun, Xijun Wang, Yan Zhang, Tony Q. S. Quek, Junyu Liu, and Jiandong Li

ABSTRACT

As a major supplement to LTE-Advanced, D2D communications underlaying cellular networks have proven efficient in offloading network infrastructures and improving network performance. The scheduling mechanism plays a key role in providing better user experience in D2D communications. However, controlled by operators, D2D communications pose specific problems that do not exist in available wireless networks. Therefore, mature scheduling mechanisms devised for cellular networks or ad hoc networks are not directly applicable to D2D communications. In this article, we first review recent research on scheduling mechanisms for D2D communications, and discuss the design considerations and implementation challenges. Then we propose an on-demand scheduling mechanism, DO-Fast, which can provide QoS differentiation capabilities. Next, we provide performance evaluations based on simulations and an experimental testbed. Finally, we conclude this article and point out possible directions for future research.

INTRODUCTION

Driven by the emerging proximity-based services, such as local advertising and social networking applications, device-to-device (D2D) communications are developed as an alternative communication method. With this technology, mobile users in proximity can establish a direct link with each other, bypassing the base stations (BSs). This offloads the network infrastructure and increases spectral efficiency. An important use case of D2D communications is public safety, which has attracted much attention in recent years. To better shape the current Long Term Evolution (LTE) for public safety, the Third Generation Partnership Project (3GPP) is committed to two main research topics: proximity service (ProSe) capability [1] and group communications system enablers (GCSEs) [2], targeting D2D communications and enhanced support for group communications, respectively. Besides public safety, other use cases of D2D communications, as specified in [1], mainly focus on network offloading and social networking applications. Based on cellular network coverage, 3GPP categorizes D2D scenarios in the following three types: in-coverage, out-of-coverage, and partial-coverage. Specifically, in the in-coverage scenario, all D2D users are covered by the BS. In the out-of-coverage scenario, no D2D users are covered by the BS. In the partial-coverage scenario, only some D2D users are in coverage.

Different from the well-known short-range communication technologies like Bluetooth and WiFi Direct, which work on the unlicensed band,[1] D2D communications are operator-controlled and share the licensed band with commercial cellular communications [3]. On the unlicensed band, interference is uncontrollable due to the lack of unified management. While operating on the licensed band, operator-controlled D2D communications can provide better user experience through effective scheduling.

Scheduling in wireless networks helps resolve problems in sharing available resources among different users. It is expected to meet network user requirements on delay, throughput, or other quality of service (QoS) measures, and to achieve specific design objectives like network throughput and user fairness. Generally speaking, scheduling mechanisms can be classified into two types: centralized and distributed. In centralized scheduling, resources are allocated by a central scheduler that is aware of global network information and all service requests. In distributed scheduling, users compete for resources in a decentralized manner and only have local knowledge of network conditions. Different from the scheduling in cellular networks or ad hoc networks, where only one type of user needs to be considered, D2D communications involve two types of users: normal cellular user equipments (CUEs) and D2D users, complicating the design of scheduling mechanisms.

The rest of the article is organized as follows. In the next section, challenges of scheduling in

Min Sheng, Hongguang Sun, Xijun Wang, Yan Zhang, Junyu Liu, and Jiandong Li are with the State Key Laboratory of ISN, Xidian University, China.

Tony Q. S. Quek is with Singapore University of Technology and Design, and the Institute for Infocomm Research, Singapore.

[1] Bluetooth operates on 2.4 GHz, while WiFi Direct can operate on both 2.4 GHz and 5 GHz.

0163-6804/15/$25.00 © 2015 IEEE

D2D communications are discussed. Following that, scheduling mechanisms in overlay mode are considered. The major design considerations and implementation challenges are first discussed, and then an overview of current research is given. Next, a distributed scheduling mechanism is proposed. Then scheduling mechanisms in underlay mode are summarized, and the differences from those in overlay mode are discussed. Following that, the proposed scheduling mechanism is evaluated based on simulations and an experimental testbed. Conclusions are given in the final section.

Challenges of Scheduling in D2D Communications

A key distinction between D2D communications and conventional ad hoc networks is that D2D communications can rely on assistance from BSs for scheduling, bringing about a better user experience. Controlled by BSs, D2D communications should be compatible with the current LTE-Advanced, which poses the following challenges in scheduling design: resource sharing and timing synchronization.

Resource Sharing

Resource sharing between D2D users and CUEs plays a key role in network performance. Two typical resource sharing modes are available: underlay and overlay. Specifically, in underlay mode D2D users and CUEs share the same channels, whereas in overlay mode D2D users have access to the channel resources that are orthogonal to those used by CUEs. In underlay mode, D2D users can fully utilize the available bandwidth to improve the data rate, while the interference from existing CUEs may interrupt a D2D user's transmission. In overlay mode, although the interference from cellular transmission is avoided, an unsuitable bandwidth partition may not satisfy a D2D user's minimum rate requirement. To determine an optimal resource sharing mode, the following factors should be considered: available bandwidth, traffic volume of CUEs and D2D users, and interference between CUEs and D2D links.

Timing Synchronization

In LTE-Advanced, scheduling for CUEs is synchronized on the timescale of one subframe (i.e., 1 ms). Asynchronous contention-based scheduling schemes used in ad hoc networks are inapplicable to D2D communications due to their incompatibility with the existing LTE architecture. By employing the fine-grained timing information from BSs (e.g., a beacon), an efficient synchronous distributed scheduler can be developed for D2D users. However, achieving exact synchronization for all D2D users is not easy, because BSs may not be completely synchronized [4]. This leads to timing offsets between D2D users at cell edges. These cell edge D2D users may fail to detect signals from each other. Moreover, in partial-coverage or out-of-coverage scenarios, where the timing signals of BSs are limited or unavailable, timing synchronization is even more challenging.

Scheduling in Overlay Mode

In overlay mode, scheduling can be centralized or distributed. With distributed scheduling, the BS may only provide access authentication and synchronization for D2D users. D2D users can page each other freely and set up connections autonomously without intervention from the BS.

In overlay mode, scheduling can be centralized or distributed. With distributed scheduling, the BS may only provide access authentication and synchronization for D2D users. D2D users can page each other freely and set up connections autonomously without intervention from the BS.

Design Considerations

Scheduling Criteria: Scheduling criteria play a key role in determining concurrent transmissions on the same frequency band. Two primary scheduling criteria are available: energy-based scheduling and signal-to-interference ratio (SIR)-based scheduling. For energy-based scheduling, the system defines an energy-based threshold to determine whether the channel can be reused by a link in the presence of existing transmissions. For SIR-based scheduling, whether links are permitted to share the same channel depends on the SIR of the received signal. As long as the imposed interference does not cause the SIR of each receiver to drop below the given SIR threshold, the new transmission is allowed to coexist with the existing transmissions. It proves that SIR-based scheduling achieves higher spatial reuse gain than energy-based scheduling [4].

QoS Requirement: In D2D communications, different D2D users may request services with different QoS requirements. Thus, it is recommended that the scheduler have QoS differentiation capabilities. Specifically, for delay-sensitive services such as voice and video conference, the scheduler should serve them preferentially to satisfy their strict delay and jitter requirements. For interactive services with strict requirements on packet loss rates, such as online games, database retrieval, and web browsing, the scheduler should allocate high-quality subchannels to them and carefully control the interference level. With regard to best effort services, such as email, SMS, and FTP, the scheduler does not need to serve them immediately due to their loose requirements on delay and data rate.

Compared to CUEs, the QoS of D2D users is much more difficult to guarantee due to the restrictions of available bandwidth and channel estimation. Due to the coexistence of CUEs, the available bandwidth for D2D users may be semi-static and affected by the cellular traffic load. In centralized scheduling, it is challenging for the BS to accurately estimate the channel quality of a D2D link, because the D2D link is set up directly and bypasses the BS. In distributed scheduling, whether the QoS of D2D users can be guaranteed also depends on the measurement of channel quality, wherein the hidden node problem needs to be considered.[2]

User Fairness: Due to different positions of D2D users and time-varying channels, absolute fairness is not realistic. In terms of timescale, fairness is classified into short-term fairness and long-term fairness. In LTE-Advanced, these two types of fairness are achieved by two corresponding schedulers: round-robin (RR) ,

[2] The hidden node problem is caused by the limited receiver sensitivity that occurs when a transmitting node is out of the sensing range of other transmitting nodes.

Technique	Scheduling criterion	QoS requirement	User fairness	CSI requirement	Algorithm	Scheduler
Opportunistic subchannel scheduling [5]	N/A, concurrent transmissions are not permitted	Guaranteeing a minimum user rate	Yes	Perfect CSI at the BS	Stochastic optimization: using Lagrangian dual and stochastic subgradient algorithms.	Centralized
Energy-efficient scheduling [6]	N/A, using adaptive modulation and coding	No QoS guarantee	No	Perfect CSI at the BS	Heuristic algorithm: allocating each RB to user with the lowest energy per bit.	Centralized
Spatial reuse [7]	Energy-based	Limiting the interference at each D2D receiver	No	Path loss and multi-path fading distributions at the BS	Greedy algorithm: allocating RBs to D2D links in slow timescale and user determines its MCS in fast timescale.	Semi-distributed
Analog signaling [4]	SIR-based	No QoS guarantee	Yes	No	Random mapping: D2D links randomly access their priorities	Distributed
Analog signaling (DO-Fast)	SIR-based	Providing QoS differentiation capabilities	Yes	Path loss and shadowing at each D2D transmitter	Group Round Robin: grouping users in terms of service type and CSI, and updating user priorities on demand.	Distributed

Table 1. Scheduling mechanisms in overlay mode.

and proportional-fair (PF). The former assigns equal available resources to users in a round-robin manner to achieve short-term fairness. The latter allocates each user a data rate that is inversely proportional to its expected resource consumption in order to provide long-term fairness as well as high network throughput. Another typical scheduler, the maximum carrier-to-interference ratio (Max C/I) scheduler, serves the user with the best channel quality in the current timeslot, causing significant unfairness to users with poor channel conditions. Compared to centralized scheduling, considering fairness in a distributed scheduling mechanism is more challenging due to the lack of global network information and a central controller.

Channel State Information: To improve the performance of D2D links, the channel state information (CSI) should be taken into account. For instance, in an orthogonal frequency-division multiplexing (OFDM) system, a user may see different channel power gains at different subchannels due to frequency-selective fading.[3] Thus, scheduling D2D users to their best quality subchannels can absolutely improve the overall system throughput.

Implementation Challenges in Distributed Scheduling

Without a central controller, distributed scheduling poses more challenges than centralized scheduling. Due to the limited receiver sensitivity, each D2D user can only acquire local information of the network, with which all D2D connections need to reach a consensus on the scheduling decision in a distributed manner.

How to Deal with the Uncertainty and Inconsistency in Topology and Network State among D2D Users: Lacking global network state, D2D users have different "views" of the network. An information exchange mechanism may help solve the problem by allowing users to collect and broadcast their local information collaboratively. However, the rapid change in network topology may incur much overhead in achieving global information. Therefore, it is recommended that distributed scheduling should not depend much on the global network state. With local information, it is of great importance to decrease collision among users and improve their consensus on the scheduling decision.

How to Implement the Scheduling Mechanism in an Efficient Distributed Way: With support from the BS, we can devise synchronized scheduling mechanisms to optimize network performance. To decrease the scheduling overhead, analog signaling [4], which is an energy-level-based signal, is preferred to digital control packets. Furthermore, user traffic characteristics and CSI should be incorporated into the scheduling mechanism with low signaling overhead.

A Survey of Scheduling Mechanisms in Overlay Mode

For centralized scheduling, the scheduler has global information of all D2D connections and the assigned resource pool. In [5], an opportunistic centralized subchannel scheduling algorithm is developed to maximize the average sum rate of the system. The BS opportunistically determines the D2D user's transmission mode and schedules each user to an exclusive subchannel by consider-

[3] OFDM is a frequency-division multiplexing scheme used as a digital multi-carrier modulation method.

ing the channel condition and QoS requirement of each D2D user. As such, the user unfairness is alleviated by satisfying a minimum rate requirement. In [6], an energy-efficient centralized scheduler is considered where the BS assigns each resource block (RB) to the D2D link with the highest ratio of transmit energy to the number of transmission bits in this RB. Although the scheduler achieves higher energy efficiency, it ignores the user QoS requirement and does not consider user fairness. Both of the algorithms in [5, 6] require perfect CSI of D2D users in each subframe, leading to very high signaling overhead. To solve this problem, a semi-distributed scheduling algorithm is proposed in [7], with which D2D links are scheduled based on the average channel power gain on a long timescale. A D2D transmitter estimates the instantaneous CSI and adaptively adjust its modulation and coding scheme (MCS) level in each subframe. However, with the energy-based criterion, the algorithm cannot fully exploit the spatial reuse gain. In addition, the QoS requirement and user fairness are not considered.

For distributed scheduling, the scheduling scheme is executed by D2D users themselves. A typical distributed scheduler is FlashLinQ [4], which is an OFDM-based synchronous medium access control/physical layer (MAC/PHY) architecture for D2D communications and works on the licensed band. To reduce the signaling overhead, an analog signaling scheme is incorporated to schedule D2D connections based on the SIR criterion. In this scheme, D2D users perform the scheduling algorithm by transmitting analog signals on different time-frequency tones. Compared to the digital control packet, which includes a sequence of bitstreams generated using a digital modulation method, analog signaling is an energy-level-based signal. Fairness is guaranteed by randomizing D2D users' access priorities over time. However, without considering traffic characteristics, FlashLinQ treats all D2D connections equally and does not have QoS differentiation capabilities. Moreover, without utilizing CSI, the random priority allocation scheme results in inferior throughput performance. Table 1 summarizes the existing scheduling schemes in overlay mode in terms of techniques, design considerations, and corresponding algorithms.[4]

Applicable Use Cases and Scenarios

Distributed scheduling mechanisms in overlay mode are applicable to public safety communications where cellular coverage may be limited or unavailable (i.e., in the partial-coverage or out-of-coverage scenario). To satisfy the stringent service requirements of high reliability and security, an overlay mode synchronous scheduler is recommended. Since public safety users are often clustered, a D2D user can be elected as the cluster head and treated as the virtual BS to assist in performing synchronization among D2D users.

Proposed Scheduling Mechanism in Overlay Mode

We propose a synchronous distributed opportunistic scheduling protocol under fairness constraint (DO-Fast) [8], aiming to provide QoS differentiation capabilities and realize on-demand scheduling for D2D connections. Moreover, we utilize CSI to improve system throughput and consider user fairness. We conduct priority-based scheduling, in which user priorities are determined by traffic characteristics and CSI.

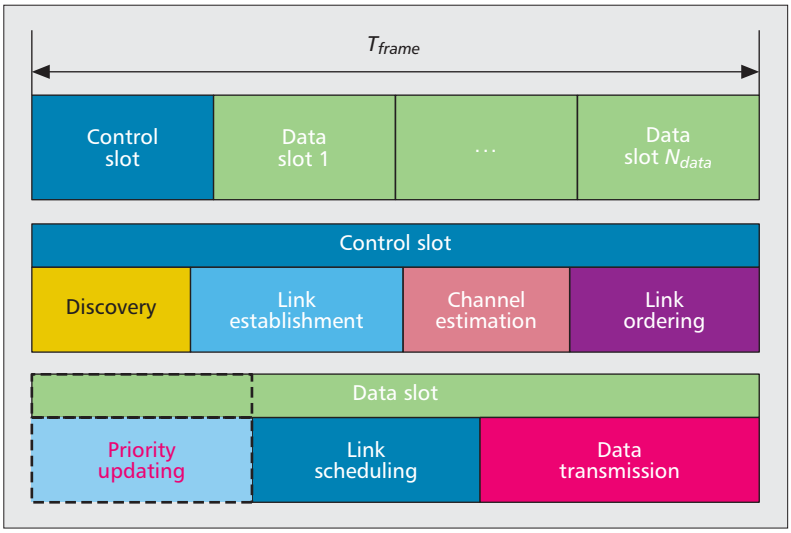


Figure 1. Frame architecture of the proposed scheduling mechanism in overlay mode.

Figure 1 shows the frame architecture, which consists of one control slot and N_{data} data slots. Timing synchronization is conducted at the beginning of the discovery phase, which can be achieved by utilizing the GPS, beacons of the BS, and so on. D2D users utilize the channel estimation phase to probe service types and channel gains of their neighbors. In the link ordering phase, each D2D user maintains a local link list by recording information collected in the previous phase. A data slot consists of priority updating, link scheduling, and data transmission. In priority updating, the user priority is updated according to a rule and causes no control overhead. The scheduling algorithm is adopted in the link scheduling phase to determine the admissible D2D links in the current time slot. Similar to [4], scheduling is implemented using an analog signaling mechanism, where D2D connections acquire distinct scheduling resources (i.e., OFDM tones) according to their priorities. Therefore, there is a one-to-one mapping from link priorities to OFDM tones, and the number of priorities is determined by the amount of scheduling resources. DO-Fast consists of the following three parts.

Traffic Report and Channel Estimation: To lower the overhead, accurate traffic characteristics or CSI is not required. We use 2 bits to distinguish three typical types of services — delay-sensitive services, interactive services, and best-effort services — and 6 bits to differentiate links with different channel qualities. We use path loss and shadowing to distinguish the users' channel qualities, and employ a QoS class identifier (QCI) to characterize the service type. For instance, the QCIs for voice, interactive gaming, and data are 1, 7, and 9, respectively.

[4] In Table 1, N/A means not applicable.

Because of the coexistence of CUEs and D2D users, the scheduler should consider the relative priorities of these two user types. If higher priority is given to cellular transmissions, the scheduling policy for D2D users is more passive. D2D users can be scheduled on condition that the existing cellular transmissions are well protected.

Local Link Ordering Scheme: The link priority is determined by using a weighted-sum method, $Priority = w_1 \cdot f(\text{traffic}) + w_2 \cdot \phi(\text{channel})$ with $w_1 + w_2 = 1$, where $f(\cdot)$ and $\phi(\cdot)$ are used to characterize traffic and channel quality, respectively. Weights w_1 and w_2 reflect the relative preference between QoS differentiation and system throughput, which can be set on demand to achieve different performance levels. Setting a higher weight for traffic provides better QoS differentiation capabilities, while a higher weight for channel quality leads to higher network throughput. Each D2D transmitter maintains a local link list with a descending priority order and obtains its scheduling resource according to the link priority. However, due to the limited sensing range, even adjacent links may obtain different local link lists. To decrease the collision probability, we design a local link ordering scheme with which each D2D connection selects its own resource index from a candidate set. For more details, please refer to [8].

Group Round-Robin Priority Updating Scheme: To improve the short-term fairness among links, a group RR (GRR) updating scheme is proposed to update the priorities of D2D connections in each data slot. In particular, D2D connections in descending priority order are divided equally into two groups: group A and group B. The two groups take turns obtaining the higher priority, and more times of higher priority are given to group A to satisfy strict requirements of these links. Moreover, within the same group links randomly select their priorities with a hash mapping algorithm in each data slot. As such, system throughput can also be improved because more scheduling opportunities are given to links with better channel quality. The ratio of higher priority time slots for group A to those for group B is set to satisfy the QoS of users with strict requirements. When there is a large number of users, it is impossible to guarantee the QoS of all users. In this case, a threshold is set to satisfy most users. A comparison of the DO-Fast scheme with the existing scheduling schemes in overlay mode is summarized in Table 1.

Scheduling in Underlay Mode

In underlay mode, D2D users share the same resources with CUEs, leading to mutual interference between CUEs and D2D users. To well coordinate the interference between these two types of users, D2D scheduling mechanisms are usually centralized and controlled by the BS. In LTE-Advanced networks, CUEs can operate in frequency-division duplex (FDD) mode or time division-duplex (TDD) mode. Thus, it is a basic requirement for operators to support D2D communications in both duplex modes.

Design Considerations

The design considerations for scheduling in overlay mode are applicable to the scheduling in underlay mode (i.e., scheduling criterion, QoS requirement, user fairness, and CSI). Moreover, the following aspects should be also taken into account.

Resource Sharing: For a given cellular transmission, the BS is the victim of interference from D2D transmitters reusing the same cellular resources in the uplink sharing. The interference depends on the D2D transmission power and channel gain between the BS and D2D transmitter. Similarly, in the downlink sharing, the receiving CUE is the victim. To schedule D2D users, operators should take into account the mutual interference between cellular transmission and D2D connection.

Relative Priority — Because of the coexistence of CUEs and D2D users, the scheduler should consider the relative priorities of these two user types. If higher priority is given to cellular transmissions, the scheduling policy for D2D users is more passive. D2D users can be scheduled on condition that the existing cellular transmissions are well protected. In this case, D2D users are treated as secondary users as in cognitive radio networks. If D2D connections have higher priority, the scheduler can make full use of resources to preferentially meet QoS requirements of D2D users. From the perspective of system performance, operators are expected to make a joint scheduling policy by considering both D2D connections and cellular transmissions to optimize the overall system. This means that no explicit priority exists, and the user making the greater contribution to the throughput or fairness tends to be scheduled first.

Implementation Challenges in Underlay Mode

On What Timescale Should the Scheduling Operate: For each scheduling, the BS needs to measure channel quality, gather users' reports, consider the available resources, and process all the information to make the scheduling decision. This leads to large control overhead and power consumption. Therefore, the timescale should not be too fine in order to ensure relatively low overhead. On the other hand, to guarantee the timeliness and accuracy of CSI, the timescale should not be too coarse. Thus, a trade-off exists between scheduling accuracy and control overhead.

What Information Should Be Reported to the BS and What Reporting Mechanisms Should Be Used: The required information is channel states of all related users. In addition, buffer status, traffic load, and service type are also helpful in scheduling design. In FDD mode, users utilize the pilot or reference signal sent by the BS to estimate the instantaneous downlink channel state and then report it to the BS through a dedicated feedback channel. The uplink channel estimation may not be as straightforward due to the lack of reference signals from users. In TDD mode, the BS can estimate the uplink channel state using the downlink pilot based on channel reciprocity. To facilitate uplink channel estimation, 3GPP defines two types of reference signals in the LTE-Advanced networks: demodulation reference signals (DMRSs) and sounding reference signals (SRSs).

Technique	Scheduling criterion	QoS requirement	User fairness	CSI requirement	Resource sharing	Relative priority	Algorithm
Interference coordination [9]	SIR-based	No QoS guarantee	No	Perfect CSI at the BS	Uplink	Prioritizing CUE	Hungarian algorithm: solving the maximum matching of a bipartite graph.
Three-stage scheduling [10]	SIR-based	Satisfying the target SINR requirement	No	Perfect CSI at the BS	Uplink	Prioritizing CUE	Kuhn-Munkres algorithm: solving the maximum weight bipartite matching problem.
Intelligent resource allocation [11]	SIR-based	Satisfying the target SINR requirement	Yes	Perfect CSI at the BS	Uplink and downlink	Prioritizing CUE	Heuristic algorithm: a D2D link with the lowest CSI of CUE-D2D receiver link shares RB with the best CSI of CUE-BS link.
Joint scheduling and resource allocation [12]	N/A, using adaptive modulation and coding	No QoS guarantee	Yes	Perfect CSI at the BS	Uplink	Prioritizing CUE	Stackelberg game: a game theory model treating CUEs as leaders and D2D users as followers.
Combinatorial auction-based scheme [13]	N/A, using adaptive modulation and coding	No QoS guarantee	No	Perfect CSI at the BS	Downlink	Prioritizing CUE	I-CA scheme: a game theory model treating CUEs as bidders and D2D links as packages.
Maximizing spatial reuse gain [14]	SIR-based	Satisfying the target SINR requirement	No	Perfect CSI at the BS	Uplink	Prioritizing CUE	GRA algorithm: using tools from graph theory and scheduling D2D links in ascending order of their interference to other D2D links.
Interference avoidance [15]	Energy-based	Limiting the interference at each D2D receiver	No	Path loss and multi-path fading distributions at the BS	Uplink	Prioritizing D2D	Distance-constrained resource sharing: controlling the distance of CUE sharing the same RB with the given D2D link.

Table 2. Scheduling mechanisms in underlay mode.

A Survey of Scheduling Mechanisms in Underlay Mode

Most prior works prioritize cellular transmissions. To reduce interference from D2D transmitters, one effective way is to limit the number of D2D links sharing the same bandwidth with existing CUEs. In [9], with the SIR-based criterion considered, an interference coordination scheme is proposed where each CUE is permitted to be reused by at most one D2D link in the uplink. D2D users probe the existing interference at each RB and report the information to the BS, with which the BS makes the scheduling decision by satisfying the CUE's QoS requirement. A three-stage scheduling scheme is proposed in [10] to maximize the network throughput while guaranteeing the QoS of both types of users. Admission control is first performed, and then power allocation is conducted for each admissible D2D link and its potential CUEs. In the last stage, a suitable uplink transmitting CUE is selected as the partner of each admissible D2D link through a maximum weight bipartite based scheme. However, both [9, 10] are proposed to maximize the total network throughput without considering user fairness.

A greedy heuristic resource allocation is considered in [11], where the CUE with the best CSI of CUE-BS link is allowed to share resources with the D2D connection with the lowest CSI of the CUE-D2D receiver link. Three schedulers are compared, RR, Max C/I, and PF, where the PF scheduler is shown to get a better trade-off between throughput and user fairness. Although the algorithm is easy to operate, it leads to large signaling overhead. In [12], a Stackelberg game framework is developed to execute joint channel allocation, power control, and user scheduling, in which a fairness coefficient is defined to adjust the trade-off between system throughput and user fairness. With perfect CSI, the BS schedules a D2D link to reuse a CUE's uplink RB based on the achievable utilities of the D2D link. However, the QoS requirements of CUEs and D2D links are not considered. The scheduling schemes proposed in [9–12] permit one CUE's RB to be reused by only one D2D link, which cannot achieve higher system throughput. To enhance the spectral efficiency, a low-complexity iterative combinatorial auctions (I-CAs)-based scheme is investigated in [13], where the system sum rate is maximized by allowing multiple D2D links to share one CUE's channel resource. However, more complicated CSI is required due to the additional mutual interference among D2D links

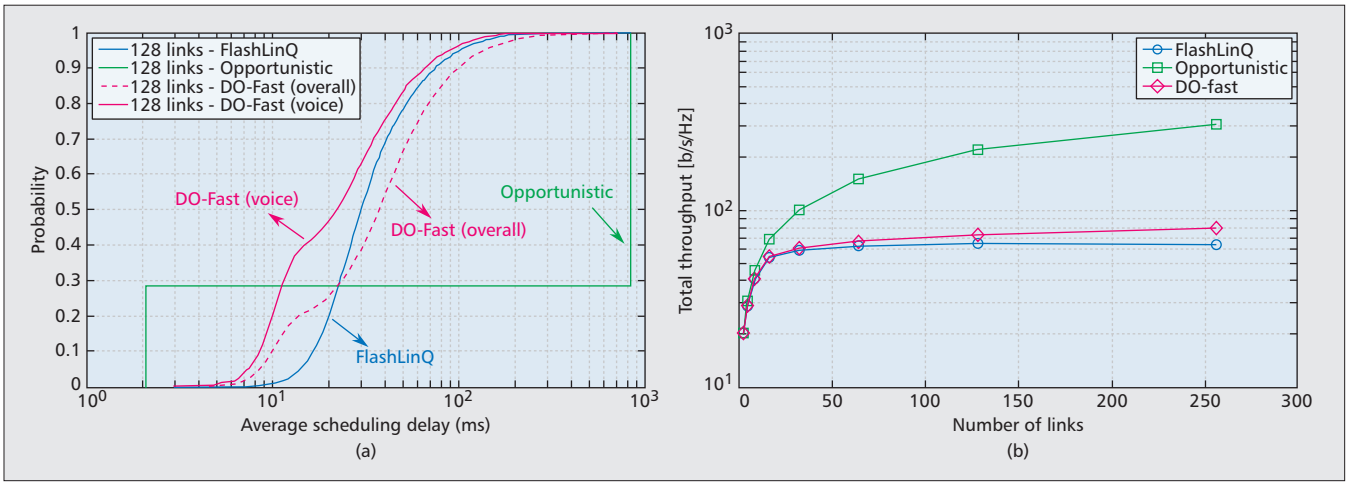


Figure 2. Simulation results: a) CDF of average scheduling delay; b) system throughput.

sharing the same resources. In addition, neither D2D user QoS requirements nor user fairness are considered. To maximize the number of admissible D2D links, [14] proposes a greedy resource allocation (GRA) algorithm by guaranteeing the QoS requirements of CUEs and D2D links. With the necessary CSI information, the BS constructs a conflict graph for all potential D2D links at each CUE's uplink RB. A D2D link imposing less interference on other D2D links will be preferentially scheduled to reuse the given RB.

D2D throughput can be improved by giving higher priority to D2D users. In [15], a D2D link is scheduled to share the cellular uplink resource by using the energy-based criterion. The interference from the CUE to the D2D receiver is controlled by keeping a minimum distance between the CUE and the D2D receiver. With the CUE's location information and the channel fading distribution, the optimal minimum distance is derived to minimize the D2D user's outage probability. However, the D2D user fairness is not involved. The existing D2D scheduling mechanisms in underlay mode are summarized in Table 2 in terms of techniques, design considerations, and corresponding algorithms.

Applicable Use Cases and Scenarios

The centralized scheduling mechanisms in underlay mode are more likely to be applied in network offloading and social networking applications in in-coverage scenarios. To efficiently offload the network infrastructure via D2D users, the BS needs to acquire the global network state and recognize the traffic hot zones. Moreover, the BS should have a good knowledge of interference between D2D users and existing CUEs. For social networking applications, a centralized scheduler is helpful to protect the privacy of D2D users and satisfy their different QoS requirements.

Performance Evaluation of Scheduling in Overlay Mode

In this section, we highlight the performance of DO-Fast in system throughput and average scheduling delay (inter-schedule delay) by comparing it with two other typical scheduling mechanisms, FlashLinQ [4] and opportunistic strategy, based on simulations and an experimental testbed, respectively. With FlashLinQ, scheduling priorities are randomized over time to maintain user fairness. Opportunistic strategy aims at maximizing the network throughput, where available resources are preferentially allocated to links with better channel quality. As shown in Table 1, the key characteristics of the DO-Fast scheduler are that it targets on-demand scheduling where users are differentiated by traffic characteristics and CSIs, and user fairness is taken into account.

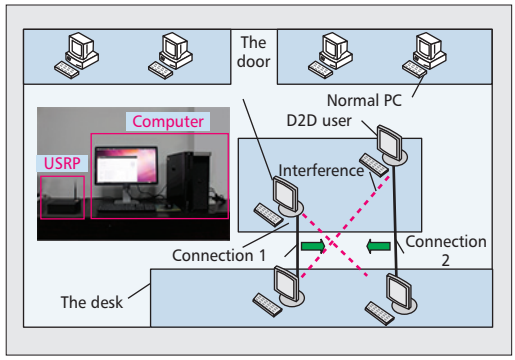


Figure 3. Testbed environment and USRP hardware platform.

We consider two types of traffic, voice and data, to represent delay-sensitive services and best-effort services, respectively. Half of the links transmit voice, and the others transmit data. In DO-Fast, group A takes the higher priority in 70 percent of the data slots, while group B takes the higher priority in the remaining 30 percent. The scheduling criterion is based on the SIR, and the decoding threshold is set to 9 dB.

Simulation Results

Figure 2a depicts the average scheduling delay. For FlashLinQ, the two types of links have the same delay performance due to lacking QoS differentiation capabilities. Opportunistic strategy just considers channel state, and only links with better channel qualities can be scheduled. This results in extreme unfairness to users with poor

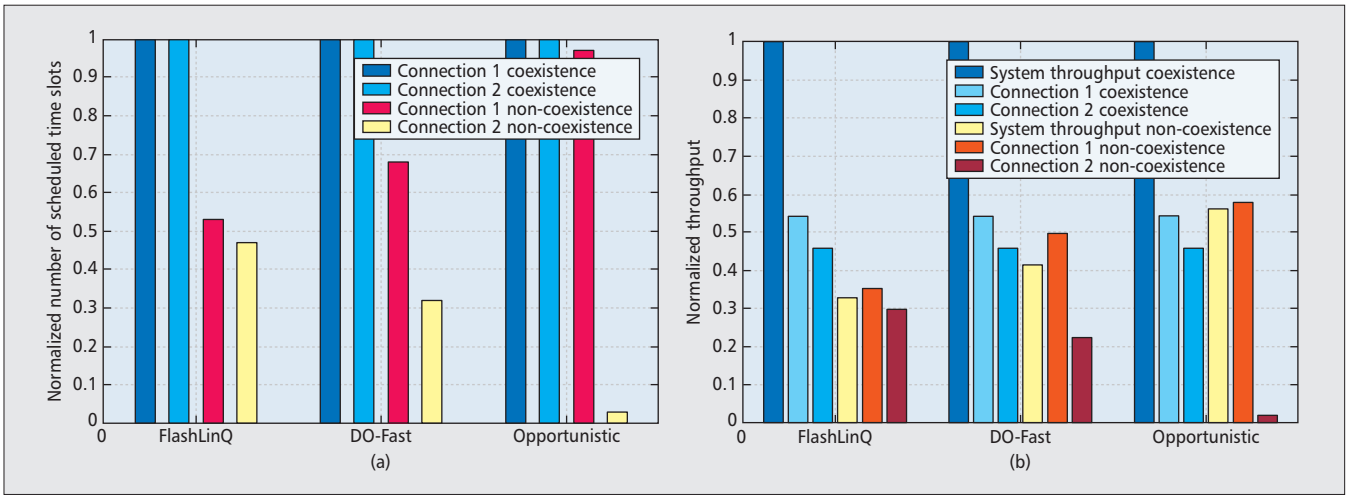


Figure 4. Experimental results: a) number of scheduled time slots; b) throughput comparison.

channels. DO-Fast provides better performance in scheduling delay for delay-sensitive services and causes little deterioration to the overall average scheduling delay. Figure 2b shows the system throughput performance. We observe that DO-Fast outperforms FlashLinQ by exploiting the CSI of links, while opportunistic strategy achieves the highest throughput. It shows that DO-Fast provides QoS differentiation capabilities and achieves higher network throughput.

Testbed Setup

With the Universal Software Radio Peripheral (USRP) board, we set up a testbed with four D2D devices forming two D2D connections, connection 1 and connection 2. As shown in Fig. 3, the testbed is placed in a 6 m × 10 m laboratory where each D2D user consists of a USRP and a PC. The two connections are differentiated by link length and service type. Connection 1 is much shorter than connection 2. The former transmits voice, and the latter transmits data. Note that connection 1 has better channel quality and stricter delay requirement than connection 2. We first separate the two connections far from each other to create a coexistence scenario. Then we move them close enough to create a non-coexistence scenario, where only the higher-priority connection can get access to the channel due to severe mutual interference. To collect results, each device records the average link rate and number of time slots to which the device is admitted or scheduled (scheduled time slots).

The experimental results are given in Fig. 4, where the average normalized number of scheduled time slots and system throughput (sum rate) are presented in Figs. 4a and 4b, respectively. Normalized number of scheduled time slots is the ratio of time slots to which a D2D link is scheduled to the total time slots, while throughput is normalized with respect to the sum rate of the two connections in the coexistence scenario. We observe that the three scheduling schemes achieve the same performance in the coexistence scenario due to simultaneous transmission of both links. In the non-coexistence scenario, the higher priority ratio between the two connections in FlashLinQ and DO-Fast is about 5:5 and 7:3, respectively. Opportunistic strategy gives nearly all access opportunities to connection 1. We observe that DO-Fast provides QoS differentiation capabilities by giving more access opportunities to services with strict QoS requirements (i.e., voice). Moreover, by exploiting CSI and considering user fairness, DO-Fast achieves higher throughput than FlashLinQ, and obtains better user fairness than opportunistic strategy.

Conclusions

In this article, we investigate scheduling mechanisms in D2D communications and provide an overview on the state of the art. We discuss technical considerations and implementation challenges of scheduling and propose an on-demand distributed synchronous scheduler, DO-Fast. This scheduler has QoS differentiation capabilities by considering the traffic characteristics, and exploits CSI to improve network throughput. A GRR priority updating scheme is further designed to enhance user fairness. Experimental and simulation results validate the superiority of the proposed mechanism.

As for scheduling, channel measurements and interference coordination are very important. Future research directions may include designing the bearer control architecture for D2D communications to help BSs collect channel reports, and combining physical layer technologies with the MAC layer to devise low signaling overhead schedulers.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61231008, 61172079, 61201141, 61301176, 91338114, and 61401320), 863 project (No.2014AA01A701), 111 Project (B08038), Basic Research Fund in Xidian University (7214497201), the Fundamental Research Funds for the Central Universities BDY011401, the SUTD-MIT IDC under Grant IDSF1200106OH, the A*STAR SERC under Grant 1224104048, and the MOE ARF Tier 2 Grant MOE2014-T2-2-002.

Future research directions may include designing the bearer control architecture for D2D communications to help BSs collect channel reports, and combining physical layer technologies with MAC layer to devise low signaling overhead schedulers.

References

[1] 3GPP TR 22.803 V12.2.0, "Technical Specification Group Services and System Aspects; Feasibility Study for Proximity Services (ProSe)," June 2013.
[2] 3GPP TS 22.468, "Group Communication System Enablers for LTE," Sept. 2013.
[3] L. Lei *et al.*, "Operator Controlled Device-to-Device Communications in LTE-Advanced Networks," *IEEE Wireless Commun.*, vol. 19, no. 3, June 2012, pp. 96–104.
[4] X. Wu *et al.*, "Flashlinq: A Synchronous Distributed Scheduler for Peer-to-Peer Ad Hoc Networks," *IEEE/ACM Trans. Networks.*, vol. 21, no. 4, 2013, pp. 1215–28.
[5] M.-H. Han, B.-G. Kim, and J.-W. Lee, "Subchannel and Transmission Mode Scheduling for D2D Communication in OFDMA Networks," *Proc. IEEE VTC-Fall*, Quebec City, Canada, Sept. 3–6, 2012, pp. 1–5.
[6] S. Mumtaz *et al.*, "Energy Efficient Interference-Aware Resource Allocation in LTE-D2D Communication," *Proc. IEEE ICC*, Sydney, Australia, June 10–14, 2014, pp. 282–87.
[7] D. H. Lee *et al.*, "Resource Allocation Scheme for Device-to-Device Communication for Maximizing Spatial Reuse," *Proc. IEEE WCNC*, Shanghai, China, Apr. 7–10, 2013, pp. 112–17.
[8] J. Liu *et al.*, "A Distributed Opportunistic Scheduling Protocol for Device-to-Device Communications," *Proc. IEEE PIMRC*, London, U.K., Sept. 8–11, 2013, pp. 1715–19.
[9] W. Zhou *et al.*, "An Interference Coordination Mechanism Based on Resource Allocation for Network Controlled Device-to-Device Communication," *Proc. IEEE/CIC ICCC*, Xi'an, China, Aug. 12–14, 2013.
[10] D. Feng et al., "Device-to-Device Communications Underlaying Cellular Networks," *IEEE Trans. Commun.*, vol. 61, no. 8, Aug. 2013, pp. 3541–51.
[11] M. Zulhasnine, C. Huang, and A. Srinivasan, "Efficient Resource Allocation for Device-to-Device Communication Underlaying LTE Network," *Proc. IEEE WiMob*, Niagara Falls, Canada, Oct. 11–13, 2010, pp. 368–75.
[12] F. Wang *et al.*, "Joint Scheduling and Resource Allocation for Device-to-Device Underlay Communication," *Proc. IEEE WCNC*, Shanghai, China, Apr. 7–10, 2013, pp. 134–39.
[13] C. Xu *et al.*, "Efficiency Resource Allocation for Device-to-Device Underlay Communication Systems: A Reverse Iterative Combinatorial Auction Based Approach," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 348–58.
[14] H. Sun et al., "Resource Allocation for Maximizing the Device-to-Device Communications Underlaying LTE-Advanced Networks," *Proc. IEEE ICCC Wksps.*, Xi'an, China, Aug. 12–14, 2013, pp. 60–64.
[15] H. Wang and X. Chu, "Distance-Constrained Resource-Sharing Criteria for Device-to-Device Communications Underlaying Cellular Networks," *IEEE Electronics Letters*, vol. 48, no. 9, Apr. 2012, pp. 528–30.

Biographies

Min Sheng (msheng@mail.xidian.edu.cn) received M. Eng and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 2000 and 2004, respectively. She has been a faculty member of the School of Telecommunications Engineering at Xidian University since 2000, where she is currently a professor with the State Key Laboratory of ISN. She was one of the New Century Excellent Talents in University by the Ministry of Education of China, and obtained the Young Teachers Award from the Fok Ying-Tong Education Foundation, China, in 2008.

Hongguang Sun (hgsun@xidian.edu.cn) received his B.S. degree from Northeastern University at Qinhuangdao, China, in 2009. He is currently working toward his Ph.D. degree with the State Key Laboratory of ISN, Xidian University. His research interests include interference management and performance analysis in heterogeneous wireless networks.

Xijun Wang (xijunwang@xidian.edu.cn) received his B.S. degree with distinction from Xidian University in 2005, and his Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in January 2012. Since 2012, he has been with the School of Telecommunications Engineering, Xidian University, where he is currently an assistant professor. His research interests include wireless communications, and cognitive radios and interference management. He was a recipient of the Best Paper Award at IEEE/CIC ICCC 2013.

Yan Zhang (yanzhang@xidian.edu.cn) received his B.S. and Ph.D. degrees from Xidian University in 2005 and 2010, respectively. He is currently an associate professor with the School of Telecommunications Engineering, Xidian University. His research interests include cooperative cognitive networks, self-organizing networks, and media access protocol design.

Tony Q. S. Quek (tonyquek@sutd.edu.sg) received B.E. and M.E. degrees from Tokyo Institute of Technology, and his Ph.D. degree in electrical engineering and computer science from MIT. Currently, he is an assistant professor with the Singapore University of Technology and Design. He was honored with the IEEE GLOBECOM 2010 Best Paper Award, the 2012 IEEE William R. Bennett Prize, the IEEE SPAWC 2013 Best Student Paper Award, and the IEEE WCSP 2014 Best Paper Award.

Junyu Liu (jyliu@stu.xidian.edu.cn) received his B.S. degree from Xidian University in 2011. He is currently pursuing his Ph.D. degree with the State Key Laboratory of ISN, Xidian University. His research interests include interference management and performance evaluation for D2D communications in wireless heterogeneous networks.

Jiandong Li (jdli@mail.xidian.edu.cn) received his B.E., M.S., and Ph.D. degrees in communications engineering from Xidian University in 1982, 1985, and 1991, respectively. He has been a faculty member of the School of Telecommunications Engineering at Xidian University since 1985, where he is currently a professor and vice director of the academic committee of State Key Laboratory of ISN. He was recognized as a Distinguished Young Researcher by NSFC and a Changjiang Scholar from the Ministry of Education, China.

AD HOC AND SENSOR NETWORKS

Edoardo Biagioni

Silvia Giordano

The goal of ad hoc networking is primarily to provide connectivity and networking where there would otherwise be none. This process has started, and such networks are beginning to appear [1, 2].

There can be many reasons for lack of connectivity. For example, in Yellowstone National Park, the single available cell tower is insufficient to provide all the bandwidth desired by visitors when the Old Faithful geyser erupts.

This particular lack of connectivity is episodic and usage-dependent, since in this case connectivity is satisfactory as long as usage is limited.

Other cases may have more structural and long-term causes for the lack of connectivity. The most typical example is that after a severe disaster, some of the infrastructure-based connectivity may fail. Many rural areas also lack connectivity. Areas with too few paying customers to justify siting expensive infrastructure are also less likely to be connected.

In situations like these and many more, ad hoc networks can and should be used for interpersonal communications. For that reason, several projects have been started to support ad hoc communication among the mobile devices that people carry throughout their daily lives. These projects include the AllNet project [3] and the SCAMPI project [4, 5]. Evaluating the effectiveness of different designs for such ad hoc networks is essential to providing good service.

The article in this series by Solmaz and Turgut provides some of the information needed for such evaluation, by characterizing mobility of people in theme parks in case of disasters. Since information about the actual movement of people in these situations is lacking, this kind of analysis is essential for providing solid evaluations.

The article looks at maps from actual theme parks and applies the Social Force Model (SFM) to effectively simulate interactions between the motions of different individuals. The 2000 individuals in the simulation are assumed to be carrying mobile devices that may interact with each other, and the article then reports some of the resulting network statistics.

As always, we want to give special thanks to the many reviewers for their careful reviews of the articles and very helpful suggestions for improvement.

REFERENCES

[1] S. Basagni *et al.*, "Mobile Ad Hoc Networking: The Cutting Edge Directions," *Elsevier Ed.*, Mar. 2013.
[2] M. Conti and S. Giordano, "Mobile Ad Hoc Networking: Milestones, Challenges, and New Research Directions", *IEEE Commun. Mag.*, Jan. 2014.
[3] E. Biagioni, " Ubiquitous Interpersonal Communication over Ad-Hoc Networks and the Internet," 47th HI Int'l. Conf. Sys. Sci., Jan. 2014, http://alnt.org/.
[4] M. Conti *et al.*, "From Opportunistic Networking to Opportunistic Computing," *IEEE Commun. Mag.*, Dec. 2010.
[5] M. Pitkänen *et al.*, "SCAMPI: Service Platform for Social Aware Mobile and Pervasive Computing," *ACM SIGCOMM Comp. Commun. Rev.*, Sept. 2012.

BIOGRAPHIES

EDOARDO BIAGIONI (esb@hawaii.edu) is an associate professor in the Department of Information and Computer Sciences at the University of Hawaii at Manoa. His research interests focus on networking, with emphasis on ubiquitous wireless networking, but have over time ranged widely from security to high-performance computing, programming languages, and human-computer interfaces. He received his Ph.D. degree from the University of North Carolina at Chapel Hill, and has been a Series Co-Editor for *IEEE Communications Magazine* since 2006.

SILVIA GIORDANO [M] (silvia.giordano@supsi.ch) received her Ph.D. from EPFL, and is a full professor at SUPSI, Switzerland, and associate researcher at CNR. She directs the NetworkingLab. She has published extensively in the areas of QoS, traffic control, and wireless and mobile networking. She is co-editor of the books *Mobile Ad Hoc Networking* (IEEE-Wiley, 2004) and *Mobile Ad Hoc Networking: The Cutting Edge Directions* (Wiley 2013). She is an ACM Distinguished Scientist, ACM Distinguished Speaker, and on the Board of ACM N2Women. She has been a Series Co-Editor for *IEEE Communications Magazine* since 2004.

Pedestrian Mobility in Theme Park Disasters

Gürkan Solmaz and Damla Turgut

ABSTRACT

Realistic mobility simulation is critical for evaluating the performance of communication networks. Although various mobility models exist, they do not capture the changes in the mobility decisions of pedestrians in specific environments. For instance, in the case of a natural or man-made disaster, the main goal is the safe evacuation of the area, creating unique pedestrian mobility patterns. In this article we focus on the scenario of evacuating a theme park in response to a disaster. We discuss the characteristics of theme parks, modeling the environment, and the mobility decisions of pedestrians. Real theme park maps are used for modeling the environment with roads, physical obstacles, and simulating disaster events. The mobility decisions of the pedestrians are based on the evacuation goal, the limited knowledge of the area, and the obstacles. The impact of the interactions between the crowd flows is modeled based on the concept of social force. The model is evaluated by comparison with the existing mobility models and the GPS traces of theme park visitors.

INTRODUCTION

Theme parks are large and crowded areas with thousands of daily visitors from all over the world. In areas such as Central Florida, which hosts five of the top ten theme parks with highest attendances in the world [1], the theme parks can represent a significant part of the region's economy. At the same time, the region also has a history of natural disasters including hurricanes, tornadoes, and tropical storms.

Managing the flow of visitors in a theme park is a significant challenge due to the high volume of visitors, as well as the large area and many physical obstacles present in the park. These challenges become even more acute in the event of a natural or man-made disaster, when a large number of visitors must be securely evacuated while performing targeted search and rescue missions at various locations in the park. To coordinate these activities, the operators of theme parks require a robust wireless communication system. As the services based on a pre-installed infrastructure might be disrupted in the case of a disaster, many recent studies considered the use of more resilient, infrastructure-independent systems. Examples of such systems include communication systems that include smartphones and other mobile devices as opportunistic communication networks [2] and wireless sensor networks with mobile sinks [3].

As these systems use network nodes whose mobility follows that of their human owners, the overall performance is dependent on the mobility of the participating humans. Thus, to evaluate the performance of the communication network, we need a realistic simulation of human mobility in the specific scenario and environment. Commonly used mobility models such as random waypoint or generic human mobility models do not approximate well human movement in theme parks, which is characterized by a highly structured, usually purposeful movement dependent on the environment. In [4] we modeled the human mobility of theme park visitors, assuming a typical day of visitors exploring the attractions of the park. In case of a disaster scenario, however, the behavior and goals of the visitors change: they try to avoid areas impacted by the disaster, find easy ways to reach secure places, and escape from the disaster areas on the fastest possible route [5].

In this article we present a mobility model of the pedestrians in theme park disasters (TP-D). We use real theme park maps to model the disaster areas that include physical obstacles, roads, and exit gates. The macro and micro mobility behaviors of the pedestrians are modeled with the theme park model and the social force concept that represents the crowd flow dynamics by social interactions. For the evaluation of TP-D, we capture the network characteristics and analyze diffusion of the mobile nodes. We analyze the model and compare its outcomes with existing mobility models and real life mobility traces of theme park visitors. Our model provides a realistic representation of human mobility and it can be used as a baseline for network simulations and the testing of disaster management strategies.

MOBILITY MODEL

CHARACTERISTICS OF THEME PARKS

Before starting to describe the model in detail, let us first explain the unique characteristics of theme parks by a mobility modeling perspective. Theme parks are very large but bounded entertainment areas. The area of a theme park includes

The authors are with the University of Central Florida.

0163-6804/15/$25.00 © 2015 IEEE

attractions and roads. We use the term attraction to denote the places in which people gather and spend time. Rides, restaurants, and live events are classified as the attractions. Roads are the pedestrian routes that connect the attractions to each other and to the exit gates. Vehicles have limited use inside theme park areas, while exit gates are usually located close to parking lots.

Theme parks are open-air areas but they also have buildings such as indoor rides, restaurants, and gift shops. The theme park areas include many man-made and natural obstacles for pedestrians. People who spend their day in theme parks have activities such as visiting rides, walking among the attractions along the roads, and eating at the restaurants.

Due to the nature of the large and crowded area, a natural or man-made disaster may have devastating effects on a theme park. As a disaster response strategy, in time of a disaster the main goal is secure and fast evacuation of the visitors. Considering an example of a tornado alert on a crowded day, the visitors should leave the park to reach the transportation services located outside the park. Since there are thousands of people leaving the park, the mobility of a single pedestrian cannot be considered independently from the others. As a result, social interactions between the pedestrians that may cause slowdowns in pedestrian flows should be considered for realistic mobility modeling.

The evacuation problem of theme parks is different from other evacuation problems. For instance, in a city scenario, the main purpose is fast evacuation of the city by the effective sharing of streets by cars and public transportation services. Other types of evacuation scenarios focus on indoor evacuation, such as evacuation from buildings or from rooms of a building. The evacuation problem of theme parks includes large areas with physical obstacles and high numbers of pedestrians. As expected, the movement of pedestrians during disasters has many differences compared with their movement in ordinary times. Because of the aforementioned characteristics, theme parks require scenario-specific mobility modeling for performance evaluation of networks as well as simulating and testing various evacuation strategies.

Modeling Theme Parks

We model the theme park as the combination of roads, obstacles, lands, and disaster events. Each road contains a set of waypoints, which are the movement points for the theme park visitors. In this case, the length of a road is equal to the sum of the distances between each pair of its consecutive waypoints. The roads show the possible routes to reach the target locations in the map. The gates are considered as the target locations and they are placed close to the borders of the park as they connect the theme park with the outside world.

Attractions are composed of man-made buildings and other structures. In ordinary times, the main goal of the visitors is to visit the attractions. In the case of a disaster, when the visitors should be evacuated from the disaster area as quickly as possible, we consider these buildings as obstacles that prevent the free flow of the visitors. Furthermore, we model the other man-made structures in the park such as fences and walls as obstacles. There are also natural obstacles such as lakes, trees, forests, rivers, and so on.

The areas which neither include the obstacles nor the roads are classified as the lands. The lands can be used by pedestrians but they are not preferred unless there are unexpected conditions on the available roads. For instance, when a road is unavailable due to an impact of the disaster event, the lands might be used instead. In some cases, lands provide shortcuts between the waypoints. Disaster areas are classified as the red-zones and they are the circular areas reflecting the effects of the disaster. In a real scenario, one can think of the red-zones as the events that damage roads or bridges, caused by an earthquake, hurricane, fire, terrorist attack, and so on. The red-zones have radius values that specify the damaged areas and active times. If a red-zone is in its active time and it affects an area including some portions of a road, the road is assumed to be unavailable at that particular time.

The model of the theme park can be created synthetically or using real maps. We use OpenStreetMap (OSM) [6] to extract the real theme park maps and parse the OSM data to generate the roads, the obstacles, the lands, and the gates. We collect the waypoints using the OSM data and connect the consecutive waypoints to create the roads. Roads have width values according to their OSM types (footway, path, and pedestrian way). Figure 1 displays an example of the real map of Epcot from Walt Disney World in Orlando (left-side), and the processed version of the map including the waypoints, the roads, the gates, and the obstacles (right-side). In this figure the small black dots represent the waypoints, while the black lines connecting the waypoints are the roads and the closed polygons are the obstacles. The model also includes red-zones that are added to the model according to their active times; however, they are not included in this initial processing of the map. While theme park models are generated computationally, it is possible to create a non-existing theme park in the design stage manually and generate a model in the same fashion.

Mobility Model of Emergency Evacuation

We describe the mobility behavior of the visitors as follows. Initially, the visitors are randomly positioned to one of the waypoints in the theme park model. Each visitor selects an exit gate among the available exit gates in the park and marks its position as the target point. A visitor is assumed to be evacuated after reaching one of the exit gates. The visitor tries to reach the target point by moving among the waypoints. Whenever the visitor reaches a waypoint, the waypoint is marked as visited. The next destination point is selected among all the visible waypoints. The visited waypoints, the waypoints positioned in a red-zone, or the waypoints that are not visible to the visitor are not taken into consideration as candidates for the next destination point. The visitor selects a new waypoint according to its distance and direction from the current position since the visitor tries to select the destination point closer to the target. The selection of the waypoints is constrained by the the visitor's knowledge about

Managing the flow of visitors in a theme park is a significant challenge due to the high volume of visitors, as well as the large area and many physical obstacles present in the park. These challenges become even more acute in the event of a natural or man-made disaster.

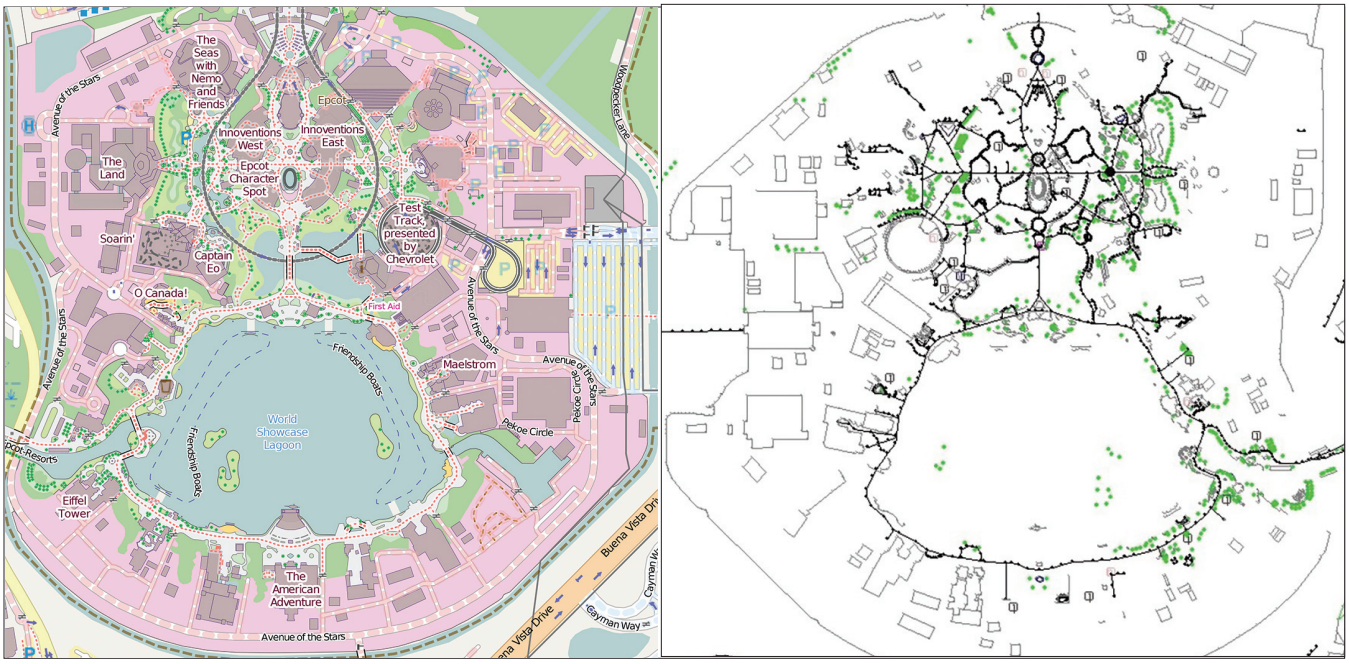


Figure 1. The maps of the Epcot park. Left: the map extracted from (OSM), right: the processed map with 1655 waypoints.

the world, the obstacles, and possible active red-zones along the way. If a visitor cannot find any waypoint as a candidate for the next destination, the new destination is selected by exploration in a random direction. The random exploration distance is a parameter that bounds the flexibility of the movement of the visitors in cases of unexpected disaster events. Another parameter that affects the flexibility is visibility. Visibility may differ according to the type and impact of the disaster. We classify all the above steps of a visitor considering the global movement starting from the initial point to the target point as the macro-mobility behavior of visitors.

The visitors have the local knowledge of their environment and they know the gate from which they entered the park. The local knowledge of the visitors is determined by the visibility parameter that specifies the visible distance for each visitor and the obstacles that may be located along the way. The visibility parameter represents the radius of the circular visible area. The visibility can have a constant or variable value according to different disaster scenarios. We mostly use a fixed value for simulations, because it basically represents the impact of the weather conditions such as heavy mist in a relatively short period of evacuation time. This model aims to serve as a baseline for testing disaster response methods. The movement of the pedestrians are modeled as they are on their own, without any help by communication devices or theme park operators for the evacuation of the pedestrians. The visitors are not assumed to communicate with each other and there is no broadcasting system for raising awareness.

The speeds of the visitors differ from each other. Each visitor has a maximum speed that depends on the physical attributes of the individual such as age, gender, and weight. The speed of each individual is determined randomly between a global minimum and a global maximum speed of the visitors. The speed of a visitor varies from 0 to the local maximum speed. The local maximum speed is the speed when the visitor is free to walk without disturbance or the obstacles. In the disaster scenario, the actual speed of a visitor is less than the local maximum speed most of the times due to the effects of social interactions.

We consider micro-mobility as the mobility of a visitor between the two consecutive waypoints separately from the macro-mobility model and the theme park model. We use the social force model (SFM) [7] which is a mathematical model of pedestrian flows sharing the same roads and used by various simulators of human mobility. According to the social force concept, the behavioral changes in human motion are actually caused by the combination of social interactions. We model the social forces on the visitors according to their social interactions with each other using SFM. By this model, the visitors adapt their speed and direction of movement from one waypoint to another. Please refer to [5] for details on how we used SFM in our application scenario.

The main effect of SFM in the theme park scenario is that the usage of the same roads by the visitors causes an increase in social interactions. This increase slows down the flow of the visitors along the roads. Since the theme parks are crowded areas with pedestrian routes, we believe that the social force model is the best-fit model to represent the crowd dynamics and the micro-mobility behavior for the evacuation of the visitors from the theme parks.

Figure 2 illustrates the complete theme park model generated using the map of Universal's Islands of Adventure park and the inclusion of the visitors and the red-zones. Forty visitors moving along the roads are represented by the yellow triangles. The shape of the triangles illustrate the directions and velocities of each of the visitors. Three red-zones are identified by the big red circles.

Model Evaluation

Simulation Setup and Metrics

The mobility metrics can be classified in three types: movement-based, link-based, and network-based metrics. The movement-based metrics are usually extracted from analyzing individuals' movement patterns. Flight lengths, average velocity, waiting times, and mean-square distances are among the movement-based metrics. The link-based metrics focus on the overall picture of the area and analyze the effects of mobility with respect to the relations between the mobile nodes. Average node density, variance of node density, average pairwise distances, and relative mobility are examples of link-based metrics. The network performance-based metrics are used to analyze the effects of mobility on the performance of the networks. In [5] we analyzed the movement-based results and evacuation times, while in this study we evaluate the link-based and network performance-based results.

The simulation of TP-D generates the mobility traces of pedestrians. We evaluate the characteristics of the resulting traces and compare them with existing synthetic mobility models as well as the GPS traces of theme park visitors. The simulations of the theme park mobility model (TP) [4], self-similar least action walk model (SLAW) [8], and Random Waypoint Model (RWP) are conducted. The TP model specifically considers realistic mobility behaviors of theme park visitors, while SLAW is a more generic human mobility model. We also include RWP since it is the most commonly used model in network simulations. The GPS traces, provided by the CRAWDAD archive, are collected from theme park visitors who spent their holidays at Walt Disney World. The traces are processed to filter out the times of traveling in a vehicle between four Disney parks.

The terrain size of the disaster simulation area depends on the size of the modeled park. We use the map of Magic Kingdom in Orlando which is the most popular theme park in the world. To model the social interactions, the circular specification of SFM is used with angular dependencies and empirical values, proposed by Helbing and Johansson [7]. Let us now summarize the simulation parameters. We have 2000 pedestrians. The simulation time is 2000 s with 0.5 s sampling time. We assume a fixed visibility of 100 m for pedestrians with 1.0 m/s maximum speed and 0.0 m/s as the minimum speed. Fifty red-zones are generated randomly with 100 m radius and 1000 s of active times. The random movement distance of pedestrians is considered 10 m. For SFM, the interaction strength is 0.11 ± 0.06 with range 0.84 ± 0.63. The relaxation time (τ) and the lambda (λ) values are set to 0.5 s and 0.1, respectively.

Analysis of the Traces

Figure 3 shows a snapshot from the simulation of 2000 pedestrians in the Magic Kingdom theme park. The pedestrian flows to the exit gates can be seen on the roads.

Experiment 1 — Average Node Degrees: The node degree of a pedestrian is defined as the number of neighbor pedestrians. The neighbors of the pedestrian are considered within the communication range of the pedestrian. In other words, two neighbors are assumed to have a wireless communication link between them if they are in communication range of each other.

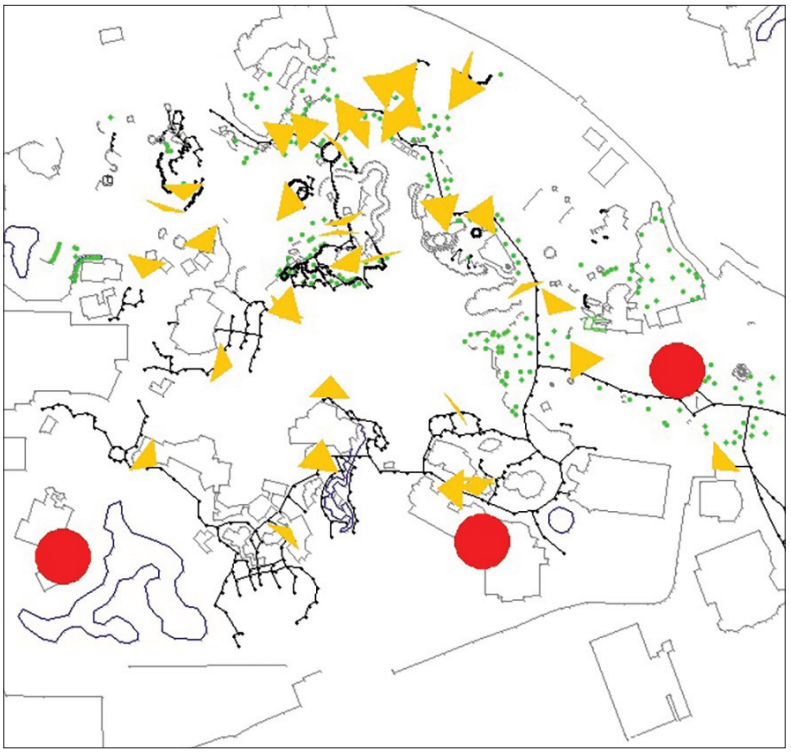


Figure 2. Islands of Adventure theme park model: 700 waypoints, 40 pedestrians and 3 red-zones.

Average node degree is a link-based metric calculated by averaging the node degrees of all the pedestrians. Basically, a higher average node degree yields better network performance. We assume a transmission range of 40 m and observe the effects of pedestrian movements on the average node degree by the varying simulation times. The results are normalized to 1000 visitors in each model.

Figure 4 shows the average node degrees by the varying simulation times for TP-D, SLAW, TP, and RWP. All the mobility models generated distinct characteristic changes in node degrees with respect to the simulation time. TP-D has the highest average node degree at the initial phase since the mobile nodes are initially distributed only on the roads while other models distribute the visitors to the entire area. We also see that the average node degrees increase steadily for TP-D, while the values may vary in short periods of time because of the shorter sampling time.

The node degrees increase much faster in the TP model due to the gathering behavior of visitors in the attractions compared to the TP-D model. In TP, visitors waiting in the queues become very close to each other. As a result, we see higher average node degrees after 2000 s. In TP-D, the pedestrians travel along the roads together, which does not produce the effect of the gathering behavior. Because of the slowdowns in pedestrian traffic, the number of people close to exit gates and main roads increase, producing the increase in the node degrees. The SLAW model has an ini-

tial phase of 500 s and the results converge to a constant level. RWP stays constant with some variances in short times caused by randomness.

Experiment 2 — Average Pairwise Distances: The distances between all pairs of mobile nodes are averaged to calculate average pairwise distances. As a link-based metric, it helps us evaluate the closeness of a node to another on average and shows the possibility to form a new network with a desired subset of all the mobile nodes. Smaller pairwise distances are expected for better network performance. As in the previous experiment, we observe the effects of mobility on the results by the varying simulation times.

As can be seen in Fig. 5, all the models again present different characteristics. TP-D has an overall constant decay of average pairwise distances. As also observed in the previous experiment, the pedestrians become closer to each other as the time moves on. An interesting difference is the fact that the significance of the effect of TP becomes weaker. This is due to the consideration of the entire population for each individual. For instance, when a visitor goes to an attraction, the pairwise distance with the other people in the same attraction becomes smaller, while the visitor's pairwise distance with people in other attractions of the park may become larger. In TP-D, on the other hand, the people move to the same target. Furthermore, since we no longer take the pedestrians, who reached the exit gates, into consideration, the fast increase due to the gathering behavior does not occur in TP-D. After the initial phase of 500 s, SLAW and RWP models have constant average pairwise distances with some variances due to randomness in the models.

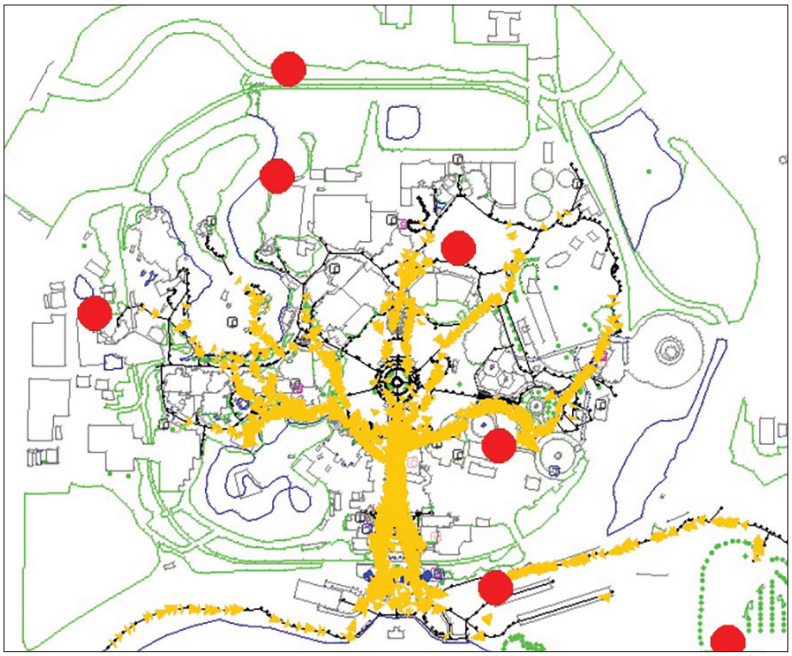


Figure 3. Crowd flows in Magic Kingdom park and effect of the red-zones.

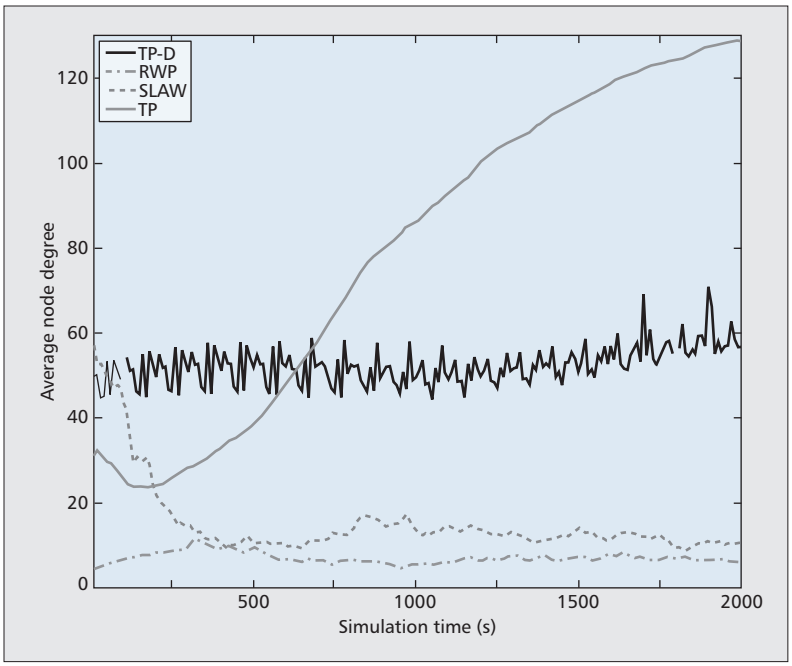


Figure 4. Average node degrees by simulation time for TP-D, TP, SLAW, and RWP.

Experiment 3 — Number of Detected Sensor Nodes: We used the mobility model as a baseline for a simulation of an opportunistic network consisting of 200 sensor nodes that are smart-phones carried by theme park pedestrians and five mobile sinks. As a disaster response strategy, mobile sinks are used for search and rescue operations and tracking pedestrians in the theme park during the evacuation process. In this experiment, the mobile sinks broadcast a message to the sensor nodes using epidemic routing and mark the sensor nodes as detected if they send acknowledgment. We observed the total number of detected sensors and analyzed the effects of various mobile sink movement strategies on the network's coverage performance. As shown in Fig. 6 physical force (PF) and road allocation (RA) based sink mobilities are the overall winners of such a scenario, reaching up to 80 percent of the nodes compared to grid allocation (GA) based, random waypoint distribution (RWD), and random target location (RTL) techniques. Finally, as expected, the experimental results reveal that increased transmission ranges allow mobile sinks to track more pedestrians in theme parks.

Related Work

In this section we summarize the related previous studies of mobility models used in network simulations. Vukadinovic *et al.* [9] propose a simple framework to simulate mobility of theme park visitors. They use OpenStreetMap for the generation of the maps and calibrate the framework parameters according to the GPS traces. While their model is a trace-based mobility model of ordinary scenarios, in this article we introduce a synthetic mobility model of pedestrians for disaster scenarios. Aschenbruck *et al.* [10] model the mobility of agents and the disaster area for crowd behavior detection. They model obstacles, dangers, and shelters as separate zones and their disaster area is divided into various sub-areas such as incident site, casualty treatment area, transport zone, and hospital zone. An event-based and role-based

mobility model for disaster areas is proposed by Nelson *et al.* [11]. In their model the movement patterns of people with different roles vary by their distinct reactions to disasters. For instance, a civilian aims to escape from a burning building while a firefighter runs into the burning building to save lives. While these studies explore different aspects of mobility modeling, none of them focus on pedestrian mobility during disasters in large-scale areas without the use of vehicles.

Conclusion

In this article we presented the TP-D model for evacuation of visitors from theme parks during disasters. The mobility of pedestrians is modeled using real maps with considerations of physical obstacles within the theme park and social interactions among the visitors. We evaluated our model in comparison with the TP, SLAW, and RWP mobility models and real-world GPS traces.

As a future work we plan to use the TP-D model as a baseline for testing disaster response strategies. One possible strategy is to use smartphones to build opportunistic social networks and broadcast messages with critical information. Moreover, we believe that the TP-D model can be adapted to various pedestrian mobility scenarios such as disasters in airports, state fairs, and shopping malls.

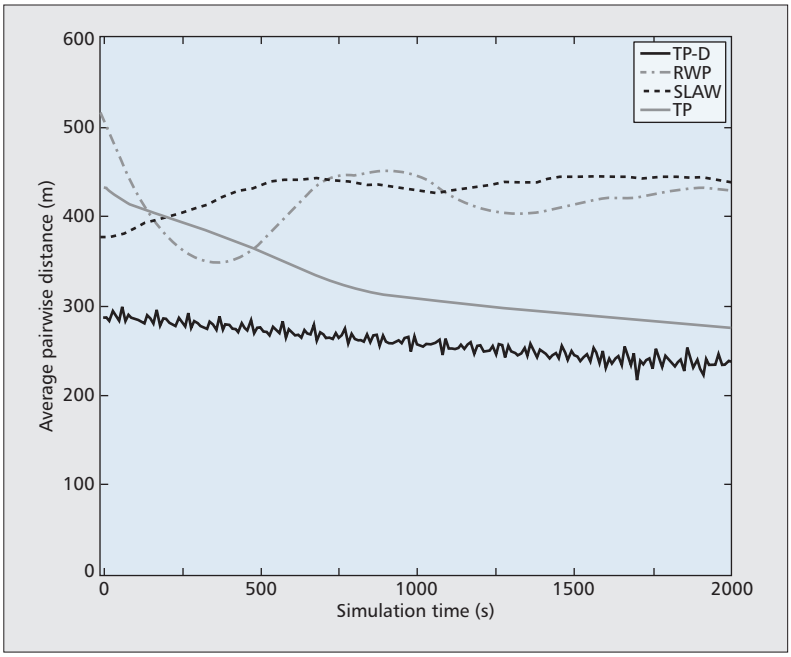


Figure 5. Average pairwise distances by simulation time for TP-D, TP, SLAW, and RWP models.

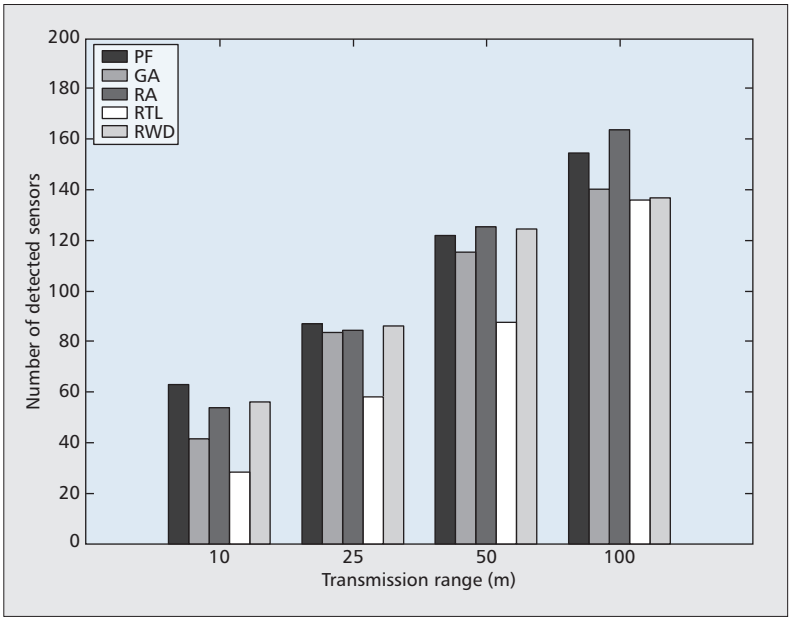


Figure 6. Number of detected sensor nodes for various mobile sink mobility models with varying transmission ranges.

References

[1] "Theme Index: Global Attractions Attendance Report," Themed Entertainment Assoc. (TEA), Tech. Rep., 2013.

[2] V. Vukadinovic and S. Mangold, "Opportunistic Wireless Communication in Theme Parks: A Study of Visitors Mobility," *Proc. ACM Workshop Challenged Netw.*, Sep. 2011, pp. 3–8.

[3] G. Solmaz and D. Turgut, "Optimizing Event Coverage in Theme Parks," *Wireless Netw.*, vol. 20, no. 6, Aug. 2014, pp. 1445–59.

[4] G. Solmaz, M. Akbas, and D. Turgut, "Modeling Visitor Movement in Theme Parks," *Proc. IEEE Local Comput. Netw.*, Oct. 2012, pp. 36–45.

[5] G. Solmaz and D. Turgut, "Theme Park Mobility in Disaster Scenarios," *Proc. IEEE Global Telecommun. Conf.*, Dec. 2013, pp. 377–82.

[6] M. Haklay and P. Weber, "OpenStreetMap: User-Generated Street Maps," *Pervasive Comput.*, vol. 7, no. 4, Dec. 2008, pp. 12–18.

[7] D. Helbing and A. Johansson, "Pedestrian, Crowd and Evacuation Dynamics," *Encyclopedia Complexity Syst. Sci.*, vol. 16, no. 4, 2010, pp. 6476–95.

[8] K. Lee *et al.*, "SLAW: Self-Similar Least-Action Human Walk," *IEEE/ACM Trans. Netw.*, vol. 20, no. 2, pp. 515–29, Apr. 2012.

[9] V. Vukadinovic, F. Dreier, and S. Mangold, "A Simple Framework to Simulate the Mobility and Activity of Theme Park Visitors," *Proc. Winter Sim. Conf.*, Dec. 2011, pp. 3248–60.

[10] N. Aschenbruck *et al.*, "Modelling Mobility in Disaster Area Scenarios," *Proc. ACM Int. Conf. Modeling, Anal. and Sim. of Wireless and Mobile Syst.*, Oct. 2007, pp. 4–12.

[11] S. C. Nelson, A. F. Harris, III, and R. Kravets, "Event-Driven, Role-Based Mobility in Disaster Recovery Networks," *Proc. ACM Workshop Challenged Netw.*, Sep. 2007, pp. 27–34.

Biographies

Gürkan Solmaz is working toward the Ph.D. degree in computer science in the Department of Electrical Engineering and Computer Science, University of Central Florida (UCF). He received his MS degree in computer science from the University of Central Florida and his BS degree in computer engineering from Middle East Technical University (METU), Turkey. His research interests include mobility modeling, mobile wireless sensor networks, and disaster resilience in networks. He is a student member of IEEE and ComSoc.

Damla Turgut is an associate professor in the Department of Electrical Engineering and Computer Science at the University of Central Florida. She received her BS, MS, and Ph.D. degrees from the Computer Science and Engineering Department at the University of Texas at Arlington. Her research interests include wireless ad hoc, sensor, underwater, and vehicular networks, as well as considerations of privacy in the Internet of Things. She is also interested in applying big data techniques for improving STEM education for women and minorities. Her recent honors and awards include being selected as an iSTEM Faculty Fellow for 2014-2015 and being featured in the UCF Women Making History series in March 2015. She was a co-recipient of the Best Paper Award at IEEE ICC 2013. Dr. Turgut serves as a member of the editorial board and of the technical program committee of ACM and IEEE journals and international conferences. She is a member of IEEE, ACM, and the Upsilon Pi Epsilon Honorary Society.

ACCEPTED FROM OPEN CALL

On the Limits of Predictability in Real-World Radio Spectrum State Dynamics: From Entropy Theory to 5G Spectrum Sharing

Guoru Ding, Jinlong Wang, Qihui Wu, Yu-Dong Yao, Rongpeng Li, Honggang Zhang, and Yulong Zou

ABSTRACT

A range of applications in cognitive radio networks, from adaptive spectrum sensing to predictive spectrum mobility and dynamic spectrum access, depend on our ability to foresee the state evolution of radio spectrum, raising a fundamental question: To what degree is radio spectrum state (RSS) predictable? In this article we explore the fundamental limits of predictability in RSS dynamics by studying the RSS evolution patterns in spectrum bands of several popular services, including TV bands, ISM bands, cellular bands, and so on. From an information theory perspective, we introduce a methodology of using statistical entropy measures and Fano inequality to quantify the degree of predictability underlying real-world spectrum measurements. Despite the apparent randomness, we find a remarkable predictability, as large as 90 percent, in real-world RSS dynamics over a number of spectrum bands for all popular services. Furthermore, we discuss the potential applications of prediction-based spectrum sharing in 5G wireless communications.

Guoru Ding, Jinlong Wang, and Qihui Wu are with PLA University of Science and Technology.

Yu-Dong Yao is with Stevens Institute of Technology.

Rongpeng Li and Honggang Zhang are with Zhejiang University.

Yulong Zou is with Nanjing University of Posts and Telecommunications.

INTRODUCTION

During the past decades we have witnessed a dramatic growth in wireless access along with the popularity of smart phones, mobile TVs, and many other wireless services. The ever-increasing demand for high data rates in the face of limited radio spectrum resources has motivated the introduction of cognitive radio (CR), which opens a potential communication paradigm to improve spectrum utilization by allowing secondary users to opportunistically access spectrum holes or white spaces unused by primary users. To enable CR, one fundamental challenge is how to reliably identify when and where spectrum holes exist.

Spectrum sensing and spectrum prediction are known as two effective enabling techniques to identify spectrum holes. Briefly, spectrum sensing determines radio spectrum state (RSS) using various signal detection methods, which have been investigated extensively in the literature (e.g. a survey in [1]). Complementarily, spectrum prediction infers unknown/unmeasured RSS from historical known/measured spectrum data by exploiting the inherent correlation and/or regularity among them, which has gained increasing attention recently (e.g. a survey in [2]). Spectrum prediction has many merits, for example, reducing sensing time and energy consumption involved in adaptive spectrum sensing [3] and increasing system throughput via prediction-based dynamic spectrum access [4], and so on.

To reap these benefits, a number of spectrum prediction techniques have been proposed, such as time series-based prediction, autoregressive model-based prediction, hidden Markov model-based prediction, neural networks-based prediction, and Bayesian inference-based prediction, etc. (e.g. the survey in [2] and the references therein). However, so far it is not clear what the upper-bound performance of various prediction techniques could be for various frequency bands. Moreover, RSS evolution patterns are generally determined by people's usage of radio spectrum. Although we rarely consider human activity in the radio domain to be totally random, current models of RSS evolution are fundamentally stochastic (see the most widely used continuous/discrete-time Markov chain models in [5]). Yet the probabilistic nature of the existing modeling framework raises fundamental questions: What is the role of randomness in RSS evolution? To what degree are RSS dynamics predictable?

This article attempts to study the interplay between the regular (and thus predictable) and the random (and thus unforeseeable) underlying real-world RSS dynamics theoreti-

0163-6804/15/$25.00 © 2015 IEEE

cally and provide certain guidance over how to apply the predicted RSS to the design of future wireless communication systems technically. Specifically, from an information theory perspective, we introduce a methodology of using statistical entropy measures and Fano inequality to quantify the degree of predictability underlying real-world spectrum measurements and provide some intuitive thoughts and conclusions. After validating the fundamental limits of predictability in RSS dynamics, this article moves forward by addressing the potential applications of the predicted RSS in 5G wireless communications.

Spectrum Data Description and Preprocessing

In order to ensure the reproducibility of the spectrum prediction analysis in this article for other researchers, we use a well known open source real-world spectrum dataset from the RWTH Aachen University spectrum measurement campaign [5]. In this article we are primarily interested in several popular services, including TV bands, ISM bands, cellular bands, and so on. The resolution bandwidth of each individual spectrum band is 200 kHz. The inter-sample time is about three minutes, which corresponds to 3360 samples in one week for each 200 kHz spectrum band.[1]

As an illustrative example, Fig. 1 shows the evolution trajectories of a one week real-world RSS, that is, measured power spectral density (PSD) values, in TV bands. Several interesting phenomena can be observed. First, the RSS dynamics for various frequency bands are significantly different; several bands are heavily loaded but others not. Moreover, randomness and regularities coexist in the RSS evolution. Very strong signals can be identified in several TV bands, and it appears that the temporal variations of signals in these bands are not as significant as those in other bands.

To further show the spectrum utilization of each 200 kHz spectrum band, Fig. 2 plots the duty cycle over the frequency under two well known detection thresholds. One threshold, 107 dBm/200 kHz, has initially been proposed in the IEEE 802.22 working group for detection of wireless microphones in 200 kHz channels in the TV bands; the other more sensitive threshold, 114 dBm/200 kHz, has been specified in the FCC's final rules [5]. As shown in Fig. 2, binary spectrum occupancy (BSO) is highly dependent on the selection of the specific detection threshold.

Until now, to characterize RSS dynamics, most of the existing studies have focused on BSO traces by analyzing the ON and OFF state evolution over time. Instead, in this article we will investigate the continuously measured PSD traces and analyze the predictability of PSD evolution over time, mainly for the following concerns: the PSD is the original raw data, while the BSO, obtained from the PSD by comparing it with a detection threshold, inevitably introduces detection or sensing errors (e.g. false alarms and missed detections) [6].

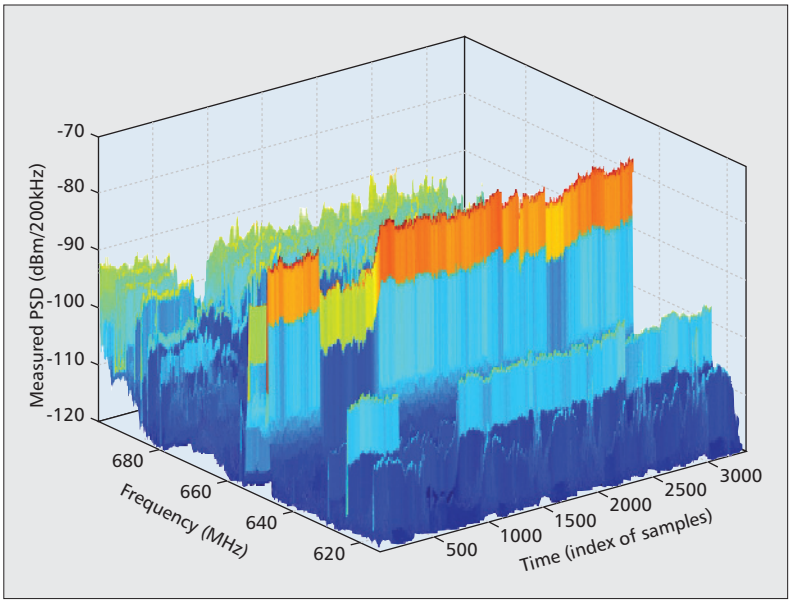


Figure 1. The 3-D view of the evolution trajectories of one week real-world RSS in the TV bands (614~698 MHz). For each 200 kHz spectrum band, about 3360 samples one week and thus 480 samples one day are plotted.

Spectrum Prediction Analysis: To What Degree is Radio Spectrum State Predictable?

In this section we first perform prediction analysis on each individual 200 kHz spectrum band separately, and then on the entire spectrum band allocated to each service statistically.

Entropy Analysis

For a given spectrum band, let X_i be a random variable representing its state at time slot i. The state of this band from time slot 1 to time slot n is a random variable series $X_1, X_2, \ldots, X_n$. Entropy is probably the most fundamental quantity characterizing the degree of predictability of a random variable series. In general, lower entropy implies higher predictability, and vice versa. Recently, entropy-based analysis has been introduced in various prediction scenarios such as atmosphere [7], network traffic [8], and human mobility [9]. The basic idea is that entropy offers a precise definition of the informational content of predictions and it is renowned for its generality due to minimal assumption on the model of the studied scenario.

Specifically, to facilitate the following entropy analysis of RSS dynamics, we first quantize the PSD values for each individual spectrum band into Q RSS levels. Then, let $S = \{X_1, X_2, \ldots, X_n\}$ denote the series or sequence of RSS levels occurred at n consecutive time slots and we have the following three entropy measures to characterize RSS dynamics:

- Random entropy $E^{\text{rand}} = \log_2 Q$, capturing the degree of predictability of the given spectrum band's evolution if each RSS level occurs with equal probability in each time slot.
- Temporal-uncorrelated entropy $E^{\text{unc}} = -\sum_{i=1}^{n} p_i \log_2 p_i$, where p_i is the probability that the i-th RSS level occurred in the sequence S. E^{unc},

[1] The original inter-sample time in [5] is about 1.8 seconds, which results in 48000 samples one day and 336000 samples one week for each individual spectrum band. To facilitate the presentation and analysis, a preprocessing procedure in this article is performed to obtain a new spectrum dataset by averaging consecutive 1000 samples.

also known as Shannon entropy or classical information theoretical entropy, is by far the most often used entropy metric, which characterizes the heterogeneity of the RSS evolution patterns without taking into account the history of the process.

• Actual entropy $E^{\text{actual}} = -\Sigma_{S_i \subset S} P(S_i)\log_2 P(S_i)$, where $P(S_i)$ is the probability of a particular time-ordered subsequence S_i occurring in the trajectory of S. Thus, E^{actual} depends not only on the occurrence frequency of each RSS level, but also the temporal order in which the RSS levels occurred, and it captures the full frequency-time structure present in a given spectrum band's revolution pattern. In practice, to calculate the actual entropy from the historical spectrum measurements, we use an estimator based on Lempel-Ziv data compression [10], which is known to rapidly converge to the actual entropy of a time series. For a time series with length n, the entropy is estimated by $E_{\text{est}}^{\text{actual}} = ((1/n)\Sigma_{i=1}^{n} \Lambda_i)^{-1}\ln n$, where Λ_i is the length of the shortest subsequence starting at the i-th time slot which does not previously appear from time slot 1 to time slot i.

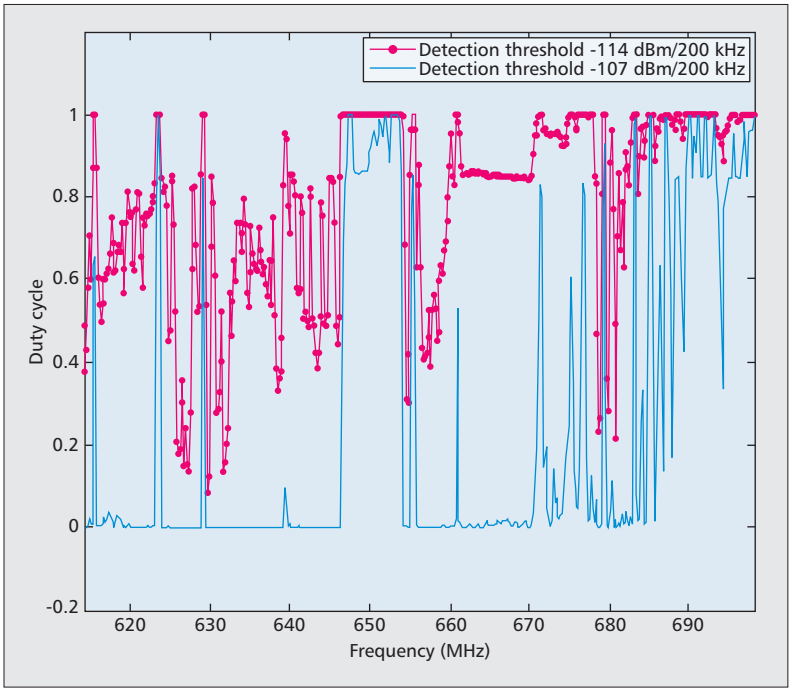


Figure 2. Impact of the detection threshold on the duty cycle in TV bands (614~698 MHz).

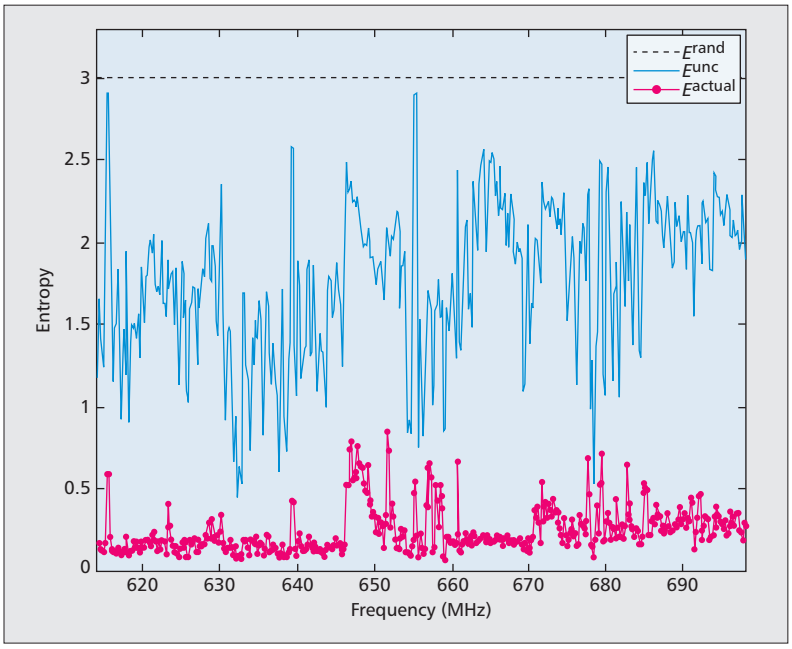


Figure 3. The entropy of the RSS dynamics in TV bands. The number of RSS levels is set as $Q = 8$ for each individual 200 kHz spectrum band and thus the random entropy $E^{\text{rand}} = \log_2(Q = 8) = 3$ bits.

Intuitively, we have $0 \le E^{\text{actual}} \le E^{\text{unc}} \le E^{\text{rand}}$, which is illustrated in Fig. 3 via analyzing the real-world spectrum measurements in TV bands. Extremely, if a spectrum band has actual entropy $E^{\text{actual}} = 0$, its RSS evolution is completely regular and thus fully predictable. However, if a spectrum band's actual entropy $E^{\text{actual}} = E^{\text{rand}} = \log_2 Q$, its trajectory is expected to follow a quite random pattern and thus we cannot predict it with an accuracy exceeding $1/Q$. As shown in Fig. 3, all spectrum bands have finite actual entropies between 0 and E^{rand}, indicating that not only a certain amount of *randomness* governs their future whereabouts, but also that there is some *regularity* in their dynamics that can be exploited for predictive purposes.

Based on the obtained actual entropy, in the following we aim to quantify the limits of the predictability of a spectrum band's next state based on its trajectory history. That is, we want to answer the question: How predictable is a spectrum band's next state given the entropy of its historical trajectory?

Predictability Analysis

An important measure of predictability is the probability Π that an appropriate predictive algorithm can correctly predict a spectrum band's future state. This quantity is subject to Fano's inequality [11]. That is, if an individual spectrum band with an actual entropy E^{actual} evolutes between Q RSS levels, its predictability $\Pi \le \Pi^{\max}$, where $\Pi^{\max}$ is determined by

$$E^{\text{actual}} = -[\Pi^{\max}\log_2\Pi^{\max} + (1 - \Pi^{\max})\log_2(1 - \Pi^{\max})] + (1 - \Pi^{\max})\log_2(Q - 1).$$

Based on this relationship, for each spectrum band, we can obtain the upper-bound predictability, $\Pi^{\max}$, through numerical calculations given Q and E^{actual}.

As an illustrative example, Fig. 4 shows the upper-bound predictability P^{max} over each 200 kHz TV band separately when the number of RSS levels is set as $Q = 8$. For comparison, the predictability of independent identical distributed (i.i.d.) Gaussian noise data with one-week samples is also plotted. We have the following observations:

• The predictability of real-world RSS data varies significantly for different spectrum bands. For example, there are a number of TV bands with the predictability higher than 0.95, which means that at most five percent of the time these spectrum bands change their states in a manner that appears to be random, and in the

remaining 95 percent of the time we can expect to predict their whereabouts. On the other hand, we also see that there are a few TV bands with predictability lower than 0.9, which means that regardless of how good our predictive algorithms are, we cannot predict with better than 90 percent accuracy the future states of these spectrum bands.

• For all TV bands, the predictability of real-world RSS data are much higher than that of the i.i.d Gaussian noise data. This demonstrates that the temporal correlation or regularity in real-world RSS data benefits the predictability.

Furthermore, from a statistical perspective, Fig. 5 shows the cumulative distribution functions (CDFs) for the predictability (with $Q = 8$) of various services, including TV bands (614∼698 MHz), ISM bands (2400.1∼2483.3 MHz), cellular bands (GSM1800 uplink 1710.2∼1784.8 MHz and GSM1800 downlink 1820.2∼1875.4 MHz), and 2.3 GHz bands (2300∼2400 MHz).[2] We have the following observations:

• Among all services, TV bands have the steepest CDF, with minimum predictability of 0.8836. Comparatively, most bands in 2.3 GHz have relatively low predictability, with the minimum close to the predictability of i.i.d. Gaussian noise data, 0.7623. ISM bands have a CDF between TV bands and cellular bands, which implies that a larger (lower) proportion of ISM bands have higher predictability than those of cellular (TV) bands.

• A predictability superiority of the GSM1800 downlink is observed over the GSM1800 uplink for spectrum bands with predictability levels in the bottom 70 percent. However, for spectrum bands with predictability levels in the top 30 percent, a predictability superiority of the GSM1800 uplink is observed over the GSM1800 downlink. That is, although a majority of the GSM1800 downlink bands have superior predictability, there are some GSM1800 uplink bands that have very high predictability levels. This somewhat conflicting observation might result from the fact that quite regular patterns of human spectrum usage exist in few GSM1800 uplink bands.

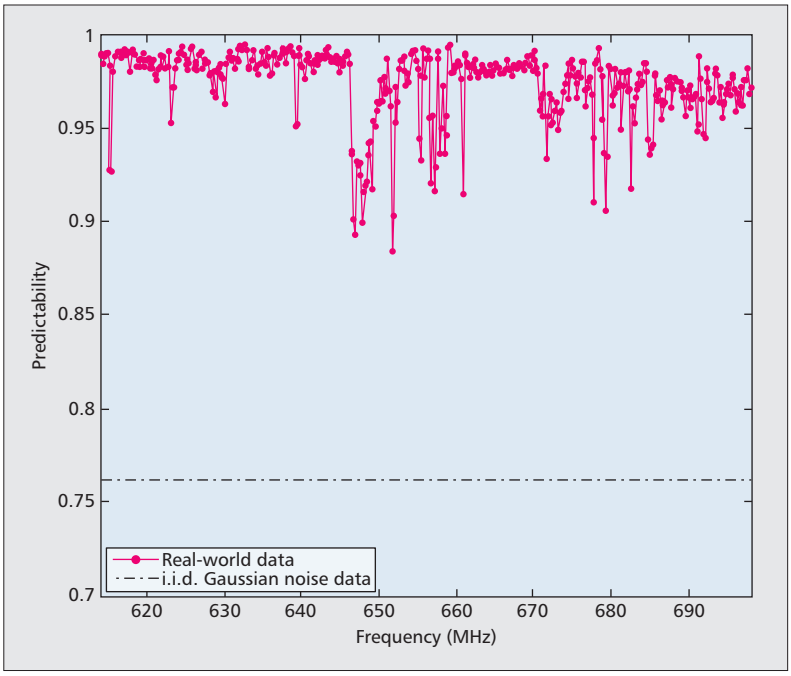


Figure 4. The predictability in RSS dynamics for TV bands.

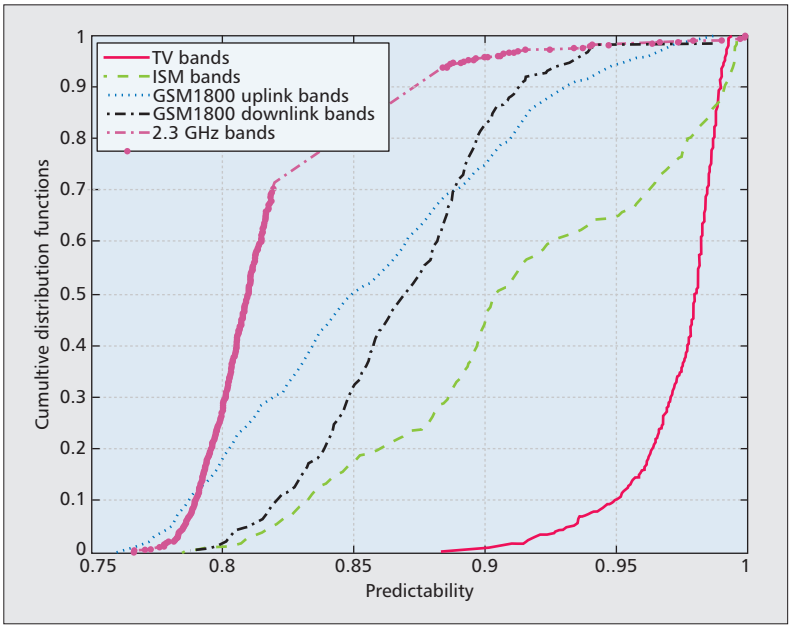


Figure 5. The cumulative distribution functions (CDFs) for the predictability of various services.

Applications: 5G Spectrum Sharing

Radio spectrum usage is an essential issue in 5G wireless communications [12]. The explosion of data rates offered by mobile Internet and the Internet of Things (IoTs) is overwhelming the allocated 2G/3G/4G radio spectrum. In the past, new cellular spectrum has typically been made available through spectrum refarming. However, clearing radio spectrum from an allocated but under-utilized usage to repurpose the spectrum band to another usage often requires many years to accomplish, which makes it difficult to keep pace with user demand of gigabit per second (Gb/s) data rates for 5G [13]. On the other hand, technological innovations such as millimeter wave communications and visible light communications can offer very high data rates; however, these disruptive technologies are mainly for small cells and low mobility usage. To provide wide area cell types, spectrum resources below 3 GHz will be needed [14].

To address these challenges, spectrum sharing is contemplated as the primary candidate, which has been well recognized as an affordable, near-term solution to meet 5G radio spectrum requirements and increase radio access network capacities for 5G content delivery. Specifically, *5G spectrum sharing is well beyond the previous studies on cognitive radio-based spectrum sharing*, since one main feature of the latter is the opportunistic primary-secondary access in unlicensed bands (such as TV white space). In contrast, as

[2] The predictability results on 2.3 GHz bands are include in Fig. 5, since Europe is currently looking at deploying licensed/authorized shared access within these bands.

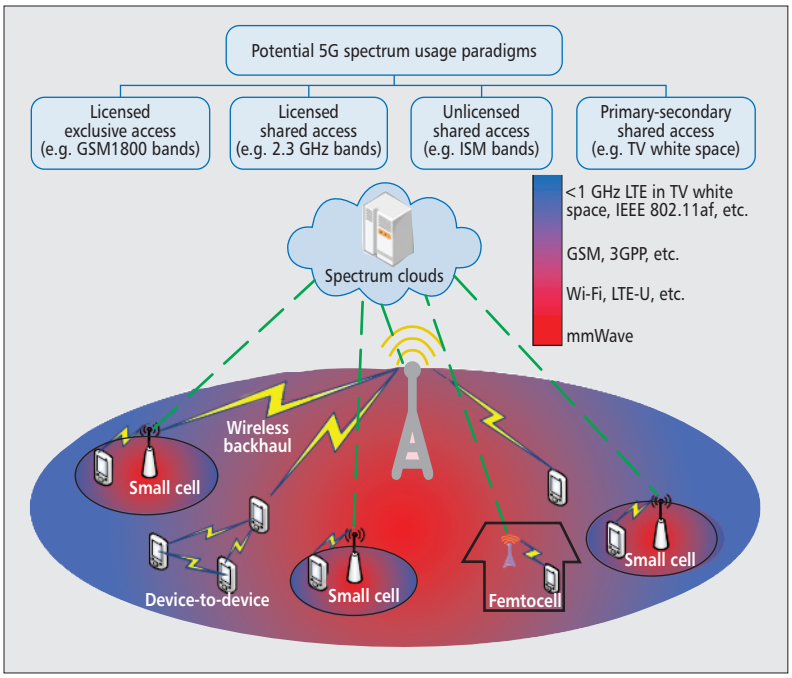


Figure 6. A vision for potential 5G spectrum usage paradigms.

shown in Fig. 6, 5G spectrum sharing may occur in both licensed bands (e.g. GSM1800 bands, 2.3 GHz bands) and unlicensed bands (e.g. ISM bands, TV bands). Moreover, one distinguishing feature of potential 5G spectrum usage is the diversity, that is, besides the licensed exclusive access in traditional cellular networks, licensed/authorized shared access, unlicensed shared access (known as LTE in unlicensed bands), and primary-secondary access will coexist [15].

Spectrum prediction will play a significant role in 5G spectrum sharing. Several potential applications are described below.

Cost-Efficient Wideband Carrier Aggregation: To meet 5G capacity requirements, it is known that no single band or air interface standard by itself fully suffices, and it is inevitable for 5G devices to aggregate the benefits of multiple (non-continuous) spectrum bands of a very wide range, possibly from several hundreds MHz bands to 30-300 GHz millimeter wave bands. Consequently, proactive schemes are expected to exploit the evolution dynamics of various spectrum bands of such a wide range, and enable wideband carrier selection and aggregation in a timely and cost-efficient manner.

Dynamic Frequency Selection and Predictive Interference Mitigation: One dominant theme for wireless evolution into 5G is network densification, which is realized mainly by increasing the density of infrastructure nodes (such as base stations and relays) in a given geographic area. It is anticipated that hyper-dense small cells are largely privately owned, and of unplanned deployment. The small cells thus need to be capable of being configured, optimized, and healed by themselves to select the communication frequency bands and not to cause any noticeable interference to the existing neighborhood networks. The knowledge from spectrum prediction can be used by the small cells to assist such autonomous processes through dynamic frequency selection and predictive interference mitigation.

Conclusion and Discussions

Predicting radio spectrum state evolution has been gaining increasing attention because of the explosive growth in demand for dynamic spectrum access. In this article statistical entropy measures and Fano inequality are exploited to quantify the degree of predictability underlying real-world spectrum measurements. The results in this article, serving as the upper-bound prediction performance, can provide a performance bound of various predictive algorithms and a general guide to the design of future wireless communication systems. Notably, it remains a challenge for the state-of-the-art prediction techniques to obtain prediction precision approaching the upper-bound predictability. Further improvement in the forecast accuracy of spectrum prediction techniques in a real-time mode are thus required.

Acknowledgment

We gratefully acknowledge the use of wireless data from Spectrum Data Archive of the Institute of Networked Systems at RWTH Aachen University. We sincerely appreciate the helpful comments and suggestions from the Guest Editor Prof. Mischa Dohler and the anonymous reviewers, who have helped us improve this article significantly. We also thank Mr. Youming Sun for his helpful input in preparing Fig. 6. This work is supported by the National Natural Science Foundation of China (Grant No. 61301160 and No. 61172062).

References

[1] E. Axell *et al.*, "Spectrum Sensing for Cognitive Radio: State-of-the-Art and Recent Advances," *IEEE Signal Process. Mag.*, vol. 29, no.3, May 2012, pp. 101–16.
[2] X. Xing *et al.*, "Spectrum Prediction in Cognitive Radio Networks," *IEEE Wireless Commun.*, vol. 20, no. 2, April 2013, pp. 90–96.
[3] G. Ning and P. Nintanavongsa, "Time Prediction Based Spectrum Usage Detection in Centralized Cognitive Radio Networks," *IEEE WCNC*, April 2012, pp. 300–05.
[4] S. Yin *et al.*, "Prediction-Based Throughput Optimization for Dynamic Spectrum Access," *IEEE Trans. Veh. Technol.*, vol. 60, no. 3, Mar. 2011, pp. 1284–89.
[5] M. Wellens, "Empirical Modelling of Spectrum Use and Evaluation of Adaptive Spectrum Sensing in Dynamic Spectrum Access Networks," Ph.D. dissertation, RWTH Aachen University, May 2010.
[6] G. Ding *et al.*, "Kernel-Based Learning for Statistical Signal Processing in Cognitive Radio Networks: Theoretical Foundations, Example Applications, and Future Directions," *IEEE Signal Process. Mag.*, vol. 30, no. 4, July 2013, pp. 126–36.
[7] R. Kleeman, A. J. Majda, and I. Timofeyev, "Quantifying Predictability in a Model with Statistical Features of the Atmosphere," *Proc. National Academy of Sciences of the United States of America*, vol. 99, no. 24, Nov. 2002, pp. 15291–96.
[8] R. Li *et al.*, "The Prediction Analysis of Cellular Radio Access Network Traffic: From Entropy Theory to Networking Practice," *IEEE Commun. Mag.*, vol. 52, no. 6, June 2014, pp. 234–40.
[9] C. Song *et al.*, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, Feb. 2010, pp. 1018–21.
[10] I. Kontoyiannis *et al.*, "Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text," *IEEE Trans. Inf. Theory*, vol. 44, vol. 3, May 1998, pp. 1319–27.

[11] R. Fano and D. Hawkins, "Transmission of Information: A Statistical Theory of Communications," *Am. J. Phys.*, vol. 29, no. 793, 1961.
[12] J. G. Andrews *et al.*, "What will 5G Be?" *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1065–82.
[13] J. Mitola *et al.*, "Accelerating 5G QoE via Public-Private Spectrum Sharing," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 77–85.
[14] Z. Khan, *et al.*, "Carrier Aggregation/Channel Bonding in Next Generation Cellular Networks: Methods and Challenges," *IEEE Network*, vol. 28, no. 6, Nov.-Dec. 2014, pp. 34–40.
[15] Special Interest Group on Cognitive Radio for 5G, "White Paper Novel Spectrum Usage Paradigms for 5G," *IEEE Technical Committee on Cognitive Networks (TCCN)*, Nov. 2014; http://cms.comsoc.org/eprise/main/SiteGen/TCCN/Content/Home/WhitePapers.html}

Biographies

Guoru Ding (dingguoru@gmail.com) received his B.S. degree in electrical engineering from Xidian University, Xi'an, China, in 2008, and his Ph.D. degree in communications and information systems from the College of Communications Engineering, PLA University of Science and Technology, Nanjing, China, in 2014. His research interests include wireless security, cognitive radio networks, machine learning, and big data analytics over wireless networks. He was a recipient of the Best Paper Awards from IEEE VTC 2014-Fall and IEEE WCSP 2009. He actively participates in the international standardization association IEEE DySPAN Standards Committee and acts as the Secretary of IEEE 1900.6 and one of the voting members both in IEEE 1900.7 and IEEE 1900.6.

Jinlong Wang (wjl543@sina.com) received his B.S. degree in wireless communications, and the M.S. degree and Ph.D. degree in communications and electronic systems from the Institute of Communications Engineering, Nanjing, China, in 1983, 1986, and 1992, respectively. He is currently a chair professor at PLA University of Science and Technology, Nanjing, China. He is also the co-chair of the IEEE Nanjing Section. He has published widely in the areas of signal processing for wireless communications and networking. His current research interests include software defined radio, cognitive radio, and green wireless communication systems.

Qihui Wu (wqhtxdk@aliyun.com) received his Ph.D. degree in communications and information systems from the Institute of Communications Engineering, Nanjing, China, in 2000. From 2003 to 2005, he was a postdoctoral research associate at Southeast University, Nanjing, China. From 2005 to 2007 he was an associate professor with PLA University of Science and Technology, Nanjing, China, where he is currently a full professor. From March 2011 to September 2011 he was an advanced visiting scholar at Stevens Institute of Technology, Hoboken, NJ, USA. His research interests include software defined radio, cognitive radio, and smart radio.

Yu-Dong Yao (yyao@stevens.edu) has been with Stevens Institute of Technology, Hoboken, NJ, USA since 2000, and is currently a professor and department director of electrical and computer engineering. From 1989 to 1990 he was at Carleton University, Ottawa, Canada, as a research associate working on mobile radio communications. From 1990 to 1994 he was with Spar Aerospace Ltd., Montreal, Canada, where he was involved in research on satellite communications. From 1994 to 2000 he was with Qualcomm Inc., San Diego, California, where he participated in research and development in wireless CDMA systems.

Rongpeng Li (lirongpeng@zju.edu.cn) is pursuing his doctorate degree in the Department of Information Science and Electronic Engineering, Zhejiang University, China. From September 2013 to December 2013 he was a visiting doctoral student in Supélec, France. From 2006 to 2010 he studied in the Honor Class, School of Telecommunications Engineering, Xidian University, China, and received his B.E. as an "Outstanding Graduate" in June, 2010. His research interests focus on green cellular networks, applications of reinforcement learning, and analysis of cellular network data.

Honggang Zhang (honggangzhang@zju.edu.cn) is a professor at Zhejiang University, China, and was the International Chair Professor of Excellence for Université Européenne de Bretagne (UEB) & Supélec, France (2012-2014). He is also an Honorary Visiting Professor at the University of York, UK. He served as the Chair of the Technical Committee on Cognitive Networks of the IEEE Communications Society from 2011 to 2012. He is currently involved in research on green communications and was the lead guest editor of the *IEEE Communications Magazine* special issues on "Green Communications."

Yulong Zou (yulong.zou@njupt.edu.cn) is a full professor and doctoral supervisor at Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China. He received a B.Eng. degree in information engineering from NUPT, Nanjing, China, in July 2006, a first Ph.D. degree in electrical engineering from Stevens Institute of Technology, Hoboken, New Jersey, USA, in May 2012, and a second Ph.D. degree in signal and information processing from NUPT, Nanjing, China, in July 2012. His research interests span a wide range of topics in wireless communications and signal processing, including cooperative communications, cognitive radio, wireless security, and energy-efficient communications. He is currently serving as an editor for *IEEE Communications Surveys & Tutorials*, *IEEE Communications Letters*, *EURASIP Journal on Advances in Signal Processing*, and *KSII Transactions on Internet and Information Systems*, among others.

ACCEPTED FROM OPEN CALL

An Evolutionary Path for the Evolved Packet System

Marc Portoles-Comeras, Josep Mangues-Bafalluy, Andrey Krendzel, Manuel Requena-Esteso, and Albert Cabellos-Aparicio

ABSTRACT

New architectural requirements appear with the evolution of mobile networks, such as the provisioning of multihoming or offloading. In general, this requires the design of ad hoc schemes on top of the EPS, which may be seen as an indicator of the need for a back-to-basics architectural analysis. This article analyzes the basic architectural principles of the EPS, and compares them with those of SDN and LISP. We then describe an evolutionary path for the EPS, by showing how, with slight modifications inspired by SDN and LISP, future mobile carrier networks could natively fulfill some of their flexibility and scalability requirements. The key design principles for that are: a generalized use of traffic flow templates (i.e., 5-tuple flows) for more flexible IP flow handling; a full decoupling of control and user plane for flexibility; and an on-demand (or pull-based) state setting at network nodes for scalability. Some examples are given to illustrate the thesis of this article.

INTRODUCTION

As future networks are developed, they will become increasingly heterogeneous and complex. This poses new requirements, such as the need for offloading at various points of the network (e.g., mobile node or femtocell) and the need for multihoming (of mobile nodes or of entire sites). In any case, scalability will be increasingly relevant as the number of network nodes increases.

As far as mobile networks are concerned, the Evolved Packet System (EPS) [1] has been selected as the carrier network for Long Term Evolution (LTE) and LTE-Advanced (LTE-A), and is being widely adopted. However, when trying to fulfill all the above requirements, EPS systematically needs to design ad hoc schemes for each of the above problems (e.g., LIPA, SIPTO, IFOM, MAPCON, or S1-flex [2–4]), which may eventually be integrated in Third Generation Partnership Project (3GPP) specifications (Releases 10–12). But additional issues appear again in subsequent releases. For instance, the NB-IFOM work item is included in Release 13 with similar objectives [5, 6]. This patchwork may be taken as a symptom that there is some design issue. We have identified three main causes of such problems: cumbersome IP flow handling, incomplete decoupling of user and control planes, and its push-oriented state information handling.

This article discusses the design principles of EPS revolving around the above three concepts and proposes slight modifications to solve the limitations of EPS when facing such new requirements. In so doing, we evolve the EPS to embed the design principles of software-defined networking (SDN) [7] and Locator Identity Separation Protocol (LISP) [8]. In fact, LISP has been easily included in SDN frameworks, such as OpenDaylight [9], as its design principles perfectly suit those of SDN. Therefore, it may sometimes be seen as an instantiation of SDN.

Despite the hype behind SDN, only recently has the Open Networking Foundation created study groups dealing with mobile and wireless networks. There have also been some previous attempts in this direction [10, 11]. However, these approaches were more disruptive in the sense that the set of protocols used was substantially modified. On the other hand, this article proposes an alternative by which most 3GPP procedures are kept, and by means of slight modifications, such novel architectural principles are introduced in the EPS. Additionally, some of the initial proponents of SDN have identified the need to move toward an SDNv2 that better considers the needs of carriers, particularly the heterogeneity of equipment (including legacy), and hence the need for a smooth migration path [12]. Bearing this in mind, this article leverages and enhances existing technologies with concepts and principles found in more disruptive approaches. As for LISP, it shares some common design principles with EPS (e.g., separation of IP address space). However, while EPS requires ad hoc mechanisms to solve some current and future needs, such as multihoming, offloading, and scalability, LISP solves them natively.

Marc Portoles-Comeras is with Cisco Systems. He was with Centre Tecnològie de Telecomunicacions de Catalunya at the time this work was done.

Andrey Krendzel is with Huawei Techonologies Oy. He was with Centre Tecnològie de Telecomunicacions de Catalunya at the time this work was done.

Josep Mangues-Bafalluy and Manuel Requena-Esteso are with the Centre Tecnològic de Telecomunicacions de Catalunya.

Albert Cabellos-Aparicio is with Universitat Politècnica de Catalunya, Barcelona, Spain.

0163-6804/15/$25.00 © 2015 IEEE

In brief, our discussion on the architectural design principles can be classified into three main groups:

- Generalized use of traffic flow templates (TFTs[1]) for more flexible IP flow handling
- Full decoupling of control and user planes for flexibility
- On-demand (or pull-based) state set up in network nodes for scalability

To illustrate this, we present some examples in which we show how, by embracing these principles, future mobile carrier networks can be simplified when fulfilling the above requirements. Moreover, adding such design principles to EPS requires significantly fewer modifications than one could imagine, making the evolutionary path toward a more future-proof mobile network easier to achieve.

On the Architectural Design of EPS, SDN, and LISP

IP Address Space Separation

A first important observation that stems from comparison of the EPS and LISP architectures is that both solutions offer a similar structure of the user plane. In particular, both architectures introduce one layer of indirection and divide the network into two address spaces. This allows the implementation and optimization of certain functions, like mobility.

Figure 1 (upper part) illustrates the EPS architecture for user plane connectivity to an IP network, called a packet data network (PDN) in 3GPP terminology, where IP addressing is separated into two different spaces, the internal and external addressing spaces.

There are network entities (PDN gateways, or PDN-GWs, and base stations, or eNBs in LTE terminology) that lay at the border of both address spaces and are able to encapsulate flows from the external IP address space to be routed through the internal address space. Furthermore, there is an internal user plane entity (Serving Gateway, or S-GW) that is able to re-encapsulate flows between user plane nodes, and can be used to optimize certain functions.

When deploying LISP to operate intra-domain we can establish a clear parallelism between the data plane structure of the network and the user plane in the EPS architecture. In particular, as Fig. 1 (lower part) illustrates, the addressing space is separated in two, creating one level of indirection between endpoint identifiers (EIDs) space and routing locators (RLOCs) space. Tunneling routers (xTRs) are the ones in charge of encapsulating (and decapsulating) flows from EID space that are routed through RLOC space. Finally, there are special elements called re-encapsulation routers (RTRs) which are used to implement and optimize certain functions that benefit from implementing middle boxes, à la S-GW in EPS.

However, despite these clear similarities in the conception of the data/user plane, the two architectures present clear differences in the way the interaction is implemented between the two address spaces. The next subsection clarifies this.

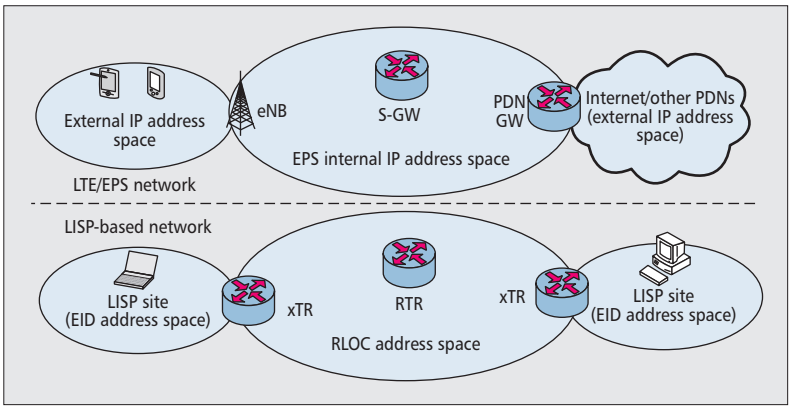


Figure 1. IP addressing architecture of LTE/EPS and LISP networks.

Embracing the IP Paradigm: IP Flows, Identifiers, and Routing Functions

EPS utilizes an all-IP network that provides better handling of end-user IP addresses and support for the whole system core (i.e., control plane and user plane) functions on top of IP-based communications. However, the EPS architecture still maintains some legacy architectural structures that limit the full adoption of the IP paradigm (i.e., the handling of flows based on TCP/IP header fields and identifiers/addresses, and the use of IP routing functions). This is especially evident in user plane elements.

First, the attach procedure [13] creates logical/virtual connections (called PDN connections) between user equipment (UE) and the PDN-GW that are only IP-aware at both ends, but not inside the network. Upon attachment, the system creates state in all the elements that form the user plane path toward the PDN-GW to ensure appropriate data forwarding. These virtual connections are defined as *bearers*, which are then used to apply appropriate quality of service (QoS) and security requirements to flows. In the uplink, the mapping between an IP flow and a bearer is done at the UE, and user's (external) IP flow information is never used again to make decisions until packets are decapsulated at the PDN-GW. The same happens with downlink flows that are only interpreted at the PDN-GW when encapsulated, and at the UE, when decapsulated.

Second, and related to the previous observation, the eNB and S-GW, in some deployment options, do not consider end-user IP information at all when performing encapsulation. Data forwarding and flow mapping to bearers, in these two entities, is done using tunnel endpoint identifier (TEID) fields and radio bearers. Upon attachment, they store enough context information to be able to encapsulate and decapsulate end-user data without the need to process external IP header information. The operation of these two entities can be understood as a natural evolution of the 3GPP architecture from the circuit switching paradigm. However, in practice, PDNs are IP-based networks [1], and IP packets generally contain enough information to be used to identify and map flows, a proof of that being the TFT filters used in the PDN-GW to map traffic and bearers, which are equivalent to the

[1] TFT is a packet filter used to identify a flow, and it is typically based on the 5-tuple (source and destination addresses, source and destination port and protocol).

The EPS signaling model used to form user plane paths is based on a push-model. That is, upon attachment of a mobile node, path information is pushed to all user plane network entities that store this information as context information.

packet matching rules of SDN forwarding nodes.

Third, it is interesting to notice that during the attachment procedure, each node of the user plane path checks, verifies, and stores multiple identifiers in relation to different network functions. At each node forming the user plane path, TEIDs are used for appropriately forwarding general packet radio service (GPRS) Tunneling Protocol (GTP) encapsulated traffic between entities. At the edges, the UE IP address assigned by the PDN-GW (along with other fields of the header of the packet) is used to differentiate incoming flows, while a number of

identifiers are used to identify the UE in different scenarios (e.g., S-TMSI is used for paging of UEs and service request, and S-RNTI is used for radio identification between the UE and the eNB [13]).

All these aspects are simplified in SDN and LISP. At the conceptual level, the former would allow flexible data plane processing by using multiple header fields (including the IP header). For instance, OpenFlow, one popular element of current SDN architectures, allows forwarding rules at nodes based on IP header fields [14]. The LISP architecture also natively handles IP-level information [8] and, in this sense, is IP-centric. This is why the LISP vs. EPS discussion is relevant here. In fact, all xTRs use end-user IP-layer information to encapsulate and forward packets through RLOC space and to implement functions that require identifiers. As a consequence of that, xTRs feature full layer 3 routing functionality and are able to differentiate flows based on an end user's IP layer data, without the need for additional information. As we show below, the possibility to differentiate IP flows in elements of the architecture highly facilitates the implementation of offloading strategies.

Decoupling of the Control and User/Data Planes

The EPS architecture aims to introduce a clear separation between control plane and user plane operation. Control plane elements, which are mostly decoupled from the user plane path, handle authentication, privacy, QoS, and mobility functions [1].

However, this decoupling between the two planes has not been developed to its full extent. In particular, when observing the control plane structure of the EPS network, one can notice that the architecture is built around the couple mobility management entity and S-GW (MME/SGW) that is located at the center of the whole system. While the MME only has control plane functionality, the S-GW shares both control and user plane responsibilities. In particular, the SGW acts as a transit point for the signaling exchange between MME and PDN-GW. Interestingly, while most control plane decisions have been decoupled from user plane elements, the responsibility to disseminate them relies on specific interfaces between the PDN-GW and SGW.

One of the founding tenets of SDN is the decoupling between control plane and data plane. Open interfaces between its building blocks are expected to bring network awareness to the applications (and vice versa) and higher flexibility through network programmability. This flexibility inherent to such full decoupling is precisely what EPS is lacking. Furthermore, what is relevant for the requirements under consideration (i.e., offloading, multihoming, and scalability) is that such flexibility is also present in nodes at the edge of the network, in which traffic diversion actually happens. This is done in SDN by applying the appropriate forwarding rules in such nodes. In a similar way, LISP xTRs (at the edge of the network) inherently provide such flexibility by potentially tunneling each flow through a different path, according to the rules configured in the mapping system (control plane).

In this way, the complexity and patchwork introduced by IFOM, MAPCON, S1-flex, or, more recently, NB-IFOM [2–5, 6] would be removed by design. The article reviews this specific case later.

Approach to Architecture Scalability: Push vs. Pull Strategies and Their Consequences

Finally, there is also a key difference between architectures in the strategy adopted to handle interfaces between control plane and user/data plane elements.

The EPS signaling model used to form user plane paths is based on a push model. That is, upon attachment of a mobile node, path information is pushed to all user plane network entities that store this information as context information. In general, a key advantage of push-based signaling models is that path changes are rapidly spread throughout the network. The downside is that network resources are (potentially) wasted by sending messages for setting up and storing state at nodes and thus used for a very low portion of time.

On the contrary, under the LISP paradigm, the interaction between data plane and control plane elements is pull-based, that is, path information is generally requested and cached only when needed. Equivalently, in certain SDN implementations (i.e., reactive flow setup), the first packet of a new flow triggers the setup of state at forwarding nodes by the control plane. This model requires the definition of some extra mechanisms to control changes in path configurations, but it scales better, as nodes only set up and store state information that they need when they need it.

Additionally, in order to ensure scalability, the EPS defines a hierarchical structure of the user plane, where the forwarding path is structured as a tree between the PDN-GW and the eNBs, acting as leaves. Thus, it inherits the rigidity of the circuit-switched world. Using this strategy, access nodes do not need to maintain much state in relation to their users, as they are assigned a default S-GW and PDN-GW to which to send data and from which to receive data. This solution to provide scalability completely removes the flexibility required to offer route optimization, as opposed to what happens with a pull-based strategy.

Feature	EPS	SDN	LISP
Embracement of the IP paradigm	Limited adoption: • per PDN-connection basis, • intermediate nodes do not use end-user IP info for packet, • multiple identifiers	Traffic processing at data plane nodes defined generically through packet matching rules, hence also including IP header fields.	Full adoption: • per IP flow basis (5-tuple), • all nodes use end-user IP info, • unique identifier
Decoupling control plane and user plane	Not full decoupling of user and control plane operation (S-GW shares user and control plane operation)	Main design principle of SDN	Separation of data and control plane functionality
Scalability	Push-based state and hierarchical user plane topology	Combination of pull- and push-based with arbitrary topologies	Pull-based state with arbitrary user (data) plane structure

Table 1. Summary of EPS design analysis in comparison with SDN- and LISP-based architectures.

The EPS architecture uses TFTs whenever an element needs to interpret external IP layer information to classify flows. TFTs are distributed during the attachment procedure with the generation of dedicated bearers, but only those elements that are going to use them store TFTs as context information.

Summary of Observations: Main Findings of EPS Design Analysis

Table 1 summarizes the main findings of EPS design analysis in comparison with SDN- and LISP-based architectures in terms of the basic design principles under discussion.

It can be seen that the EPS presents limited adoption of these key design principles. As a consequence, the EPS faces a series of problems when dealing with more flexible data networking requirements, which are hence handled through various patches (e.g., to implement offloading or multihoming). In the following three sections, we discuss current challenges caused by these limitations. We then propose an evolutionary path for EPS in the sense that with small modifications to 3GPP procedures, the EPS can embrace some of the key SDN and LISP design principles, hence solving by design what otherwise is solved through ad hoc patches.

Flexible Flow Handling: The Challenge of Offloading Traffic

The Challenge of Offloading Traffic

Mobile data offloading has become one of the key strategies adopted to face the exponential growth of mobile data in cellular networks. 3GPP approached the problem with proposals like LIPA and SIPTO [15].

LIPA is an offloading technique by means of which a UE through a base station is able to exchange data with IP-capable entities within its local network. SIPTO is an offloading technique by means of which the mobile operator is able to offload certain types of traffic through a network node close to the UE's point of attachment.

This subsection builds on the observation that LIPA and SIPTO are, in fact, solutions that introduce external IP address awareness in elements of the user plane path that do not normally use this information (e.g., eNB). However, as they are currently defined in [15], LIPA and SIPTO developed mechanisms to circumvent the requirement to use external IP addresses with support of multiple PDN connections. In a scenario with full embracement of the IP paradigm in all elements of the architecture, LIPA and SIPTO would reduce to configuring routing tables and rules.

Proposed Solution for the Challenge

The EPS architecture uses TFTs whenever an element needs to interpret external IP layer information to classify flows. TFTs are distributed during the attachment procedure with the generation of dedicated bearers, but only those elements that are going to use them (the PDN-GW, the UE, and, in some scenarios, the S-GW) store TFTs as context information. In fact, such TFTs could be seen as equivalent to the packet matching rules one may find in data plane nodes of SDN.

Interestingly enough, transitioning from the current EPS architecture to fully embracing the IP paradigm requires minimal changes to the EPS architecture. Indeed, the proposal here is to extend the use of TFT filters to all the elements in the user plane path. This includes storing TFT information in eNBs and S-GWs in addition to the PDN-GW, which was already using this information. Even more, an analysis of the attachment procedure reveals that during this process enough TFT-related information is carried through all interfaces that it can be exploited and stored locally at each of the user plane elements. As a result, eNBs and S-GWs just need to store this information that was previously ignored.

In order to prove the previous statement, we have analyzed the possibility of using TFT filters in all nodes of the user plane path to map flows to bearers in [16]. The study provides a detailed analysis revealing how none of the interfaces of the EPS architecture needs to be neither modified nor extended to support the dissemination of TFT information to all user plane entities. The flow diagram illustrating the TFT dissemination process specified by 3GPP is shown in Fig. 2. The interested reader will find a detailed description of all the necessary steps in [16].

Offloading with a TFT-Based Architecture

With an EPS system embracing the IP paradigm (or generically, an arbitrary header field matching rule), as described above, the problem of offloading is reduced to a question of distributing appropriate TFTs and routing information to affected nodes from the centralized control we discuss in the following section.

In particular, under the proposed scheme, the process of offloading traffic in a SIPTO scenario would not differ much from the process of establishing a dedicated bearer, where the eNB receives a TFT filter associated with the traffic that needs to be offloaded, as well as the alternative (offloading) path to use. It can be noticed here that, once using TFTs, the base station is able to differentiate particular flows and has the possibility to implement SIPTO functionality without requiring that the UE support multi-PDN connectivity, hence making it transparent to the UE.

In the particular case of LIPA, we need to ensure that TFT rules are applied before routing rules [14] to ensure that traffic not meant to be offloaded (i.e., traffic that must traverse the core network) is not routed locally before being tunneled to the core.

Decoupling Control and User/Data Plane Functionality: The Challenge of Multihoming

Multihoming

The multihoming problem appears repeatedly in the EPS architecture in various flavors and applied at different network entities. In the IFOM (and NB-IFOM in 3GPP Release 13) scenario, multihoming is required so that the end user is able to use more than one access network to send data and the PDN-GW can also send flows through different paths through a single PDN connection. In the MAPCON scenario, multihoming is required so that the UE can use more than one PDN connection through 3GPP and non-3GPP access technologies. Another initiative that may require the configuration of multihomed entities is the introduction of the S1-flex, where the system introduces the mobility management entity (MME) (and its associated S-GW) redundancy. The S1-flex proposal only considers the possibility of having a setup of the type active/multiple-passive S-GW. With this setup, only one S-GW can be active at a given moment of time, and this is the only one the eNB can use to forward data toward the PDNGW.

In all these cases, multihomed elements of the network core face the challenge of also having to multihome their control plane procedures. As a consequence, duplicating user plane functions leads to having to solve a problem in the coordination of control plane functions.

The limited decoupling between the control and user planes in EPS entails additional challenges in multihoming scenarios. However, with fully decoupled control and data planes, as in SDN, multihoming user plane elements (e.g., for load balancing) is not an issue.

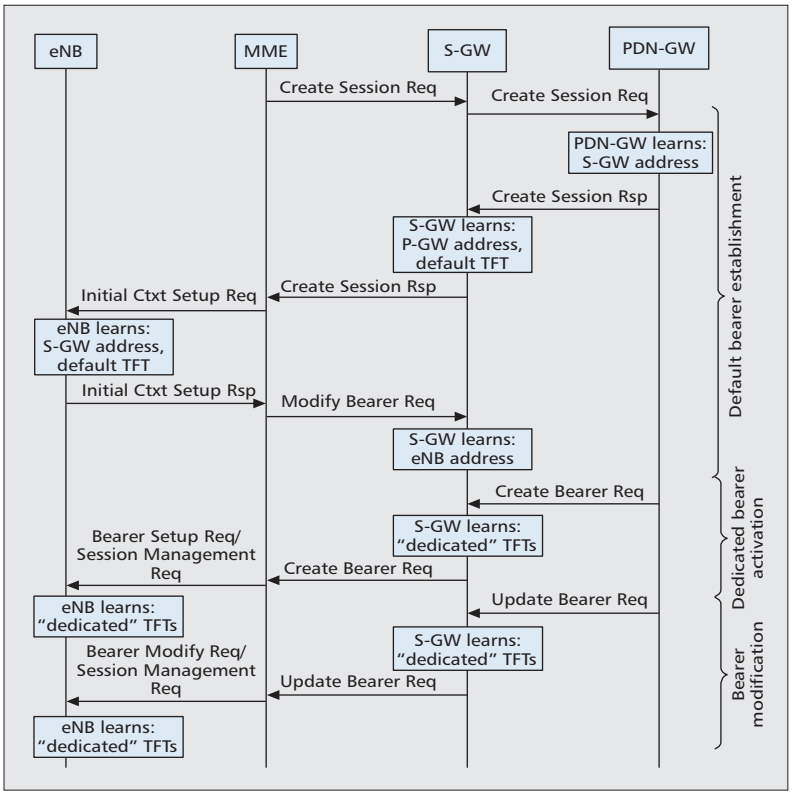


Figure 2. Dissemination of TFTs between all user plane entities.

Proposed Alternative for the Architecture: Full Decoupling of Control and User Plane

Since the MME is the central control entity of the EPS, we propose to move all control plane responsibilities of the S-GW to the MME, given the limitations introduced by having control plane responsibilities in user plane elements (e.g., for multihoming). As a consequence, the MME becomes the core responsible for the establishment and maintenance of the user plane path during the attach process and during mobility events.

Therefore, one additional interface must be established between the MME and the PDNGW, which plays the role of S5/S8 interface for control plane information (Fig. 3). Additionally, the S11 interface (between the MME and the SGW) needs to be extended to accommodate those signaling messages that are directly exchanged between the PDN-GW and the SGW.

In [16], we have analyzed the feasibility to move the messages exchanged between the PDN-GW and the MME in the current EPS architecture to the new proposed interface. The study reveals that both the attachment and the mobility procedures can be supported without modification using the new proposed interfaces and with some extension of the functionality provided through the S11 interface. For instance, in the considered procedures [16], out of 10 signaling messages currently exchanged through a combination of S11 and S5 interfaces, 8 are sent through the new interface, and 2 additional messages are sent through S11. As a result, the S5 interface is released from control plane operations.

Notice that this article focuses on connectivity and mobility management, given its focus on fundamental design principles underlying the EPS. QoS and security should be given adequate treatment in future refinements, given their importance.

Multihoming with Full Decoupling of User Plane and Control Plane Functions.

Assuming SDN-like full decoupling of control and user/data planes as proposed above, implementing multihoming in the user plane path is reduced to informing the different elements of the multiple options they have to forward data.

An example of this would be the implementation of an advanced S1-flex scenario where the eNB receives a list of alternative S-GWs to use simultaneously. It could be done in two ways. First, the MME can send to the eNB a signaling message (e.g., Initial Context Setup Request/Attach Accept [13]) including addresses of several candidate S-GWs for the user plane instead of one, as is currently done. Second, the MME can send to the eNB the same signaling message, but several times, with a different S-GW address in each. Similarly, the PDN-GW can receive the same information about candidate S-GWs directly from the MME responsible for this particular user plane path.

The same approach can be followed to implement NB-IFOM and MAPCON scenarios [16], where the MME informs the corresponding nodes about the multiple forwarding options to send and receive information.

Approach to Network Scalability: The Challenge of Route Optimization

Network Architecture and Route Optimization

The logical architecture of the user plane of an EPS system is highly hierarchical, with the PDN-GW being the root of the system and the eNBs located at the leaves. The advantage of this architecture is that it scales when the number of users and eNBs grow without increasing the complexity and requirements of eNBs. The reason is that they just need routing information to reach their corresponding S-GW and PDN-GW, and are completely unaware of other S-GWs and eNBs. However, for this same reason, it is hard to offer the flexibility required by novel scenarios.

More specifically, route optimization (Fig. 4) in this architecture is challenging. For example, when two UEs belonging to the same network are communicating, the flow must traverse the complete hierarchy (up to the PDN-GW) and back to the correspondent UE. Providing direct communications (i.e., route optimization) between eNBs in the current push-based EPS would require that all elements receive and store information about all the rest of elements, and this would not scale.

Alternatively, in a pull strategy, nodes just request the information they need and do not cache information on inactive communications.

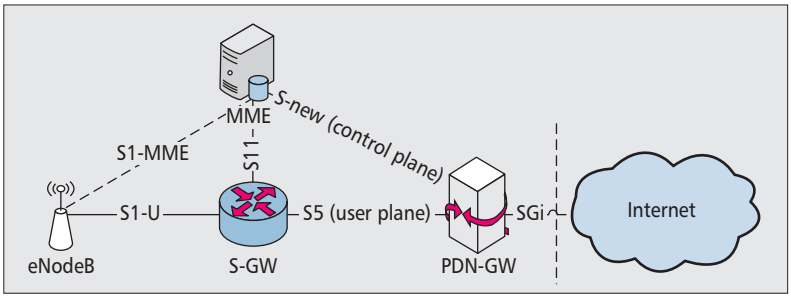


Figure 3. MME with interface to every node.

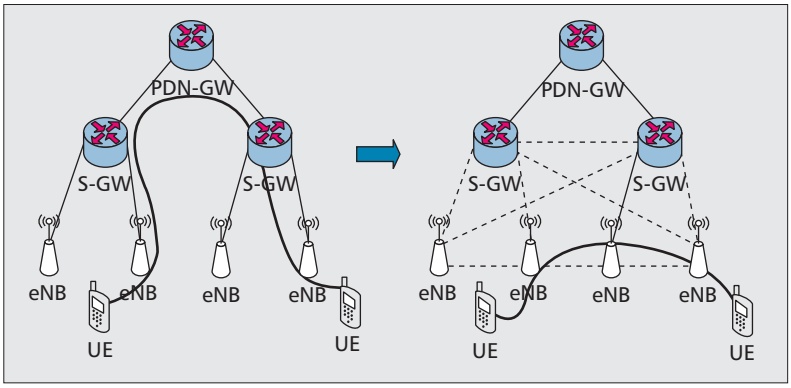


Figure 4. Example of route optimization in an EPS network.

In this way, flexibility and scalability can be simultaneously achieved.

Proposed Alternative for the Architecture: From Push to Pull

Following the reasoning above, our proposal is to move from a push-based scheme to disseminate information to a pull-based scheme, where eNBs request information about flows when they require it.

More specifically, the attach procedure would be used to establish the default bearer following regular 3GPP procedures. After that, when the UE generates the first uplink packet of a new flow, the eNB would request (i.e., pull-based) whether a dedicated bearer must be established to support the flow. Such dedicated bearers would not be constrained by the hierarchical architecture and could be established between any two elements of the system (e.g., two eNBs) thanks to the data plane information provided by the control plane. We have analyzed the requirements to use a pull-based strategy to support path formation in EPS systems in [15]. Slight modifications are needed in some EPS procedures, but the basic structure of the procedure can be maintained. For instance, the Initial Context Setup Request/Attach Accept message [13] must contain the destination eNB address.

Scalability and Route Optimization with a Pull Scheme

With the scheme proposed, the problem of route optimization is moved to the control plane, which decides, on a per-request basis, whether a flow must follow the hierarchy up to the PDN-

Challenge	3GPP initiatives	EPS limitations	Proposed solution
Traffic offloading	LIPA, SIPTO	External IP address unawareness in some of the nodes of the user plane; UE needs to support multi-PDN connectivity	Distributing TFTs and routing information to all nodes of the user plane to differentiate IP flows
Multi-homing	IFOM, MAPCON,S1-flex, NB-IFOM	Duplicating user plane operation in coordination with control plane functionality is an issue because of limited decoupling between planes	Direct interface between MME and PDN-GW for full decoupling of user and control operation
Route optimization	X2, LIPA, SIPTO	Push-based approach for path formation and hierarchical user plane topology (forwarding always through the PDN-GW)	Pull-based approach on a per-request basis and arbitrary topology for data path formation

Table 2. Network challenges, related 3GPP concepts, EPS limitations, and proposed solutions.

GW or can be routed directly between nodes of the topology.

As an example, during a handover, X2-based data forwarding would be simplified. When the source eNB receives data for a node that was no longer under its control, it would request the establishment of a dedicated bearer to forward the data and receive information about the destination eNB. As a result, data would be directly encapsulated between the two eNBs without the need for complex logical link setups.

We focused our above discussion on the route optimization problem among two nodes of the same network. However, the same pull-based approach could be used for LIPA and SIPTO toward local or external nodes, hence benefitting from the same advantages. In this case, offloading rules would not need to be preconfigured in advance, but when the UE actually starts sending data of a given flow, which allows the control plane to dynamically adapt to network conditions in a scalable way.

Summary

This article presents an evolutionary path for EPS toward a more flexible and future-proof architecture to fulfill current and future requirements (e.g., multihoming, offloading, route optimization). Such architectural reasoning exploits design principles that have shown their potential in the SDN and LISP contexts. Our proposal is evolutionary in the sense that the minimum possible subset of modifications to current 3GPP procedures is applied to include the new design principles in EPS.

This article claims that the flexibility and scalability required by novel scenarios can be provided:

- By using TFTs at each data plane node for more flexible IP flow handling (similar to SDN packet matching rules or LISP packet handling at xTRs)
- By introducing a new interface between the MME and the PDN-GW for full decoupling of user and control planes à la SDN
- By applying a pull-based model for on-demand state setup at network nodes for efficient path handling

With the proposed modifications, the above requirements could be natively solved by EPS without the need for ad hoc schemes.

The current data networking challenges, 3GPP initiatives, EPS limitations, and proposed solutions to overcome them are summarized in Table 2.

Acknowledgments

This work has been partially funded by the Spanish Ministry of Science and Innovation under grant TEC2011-29700-C02-01 and TEC2011-29700-C02-02. The authors would also like to acknowledge the valuable comments of Scott Mansfield and the help of Huub van Helvoort.

References

[1] M. Olsson *et al.*, *SAE and the Evolved Packet Core: Driving the Mobile Broadband Revolution*, Academic Press, 2009.
[2] F. Rebecchi *et al.*, "Data Offloading Techniques in Cellular Networks: A Survey," Accepted for publication in *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 2, 2nd qtr. 2015.
[3] Small Cell Forum, "Integrated Femto-WiFi (IFW) Networks," SCF Release 4, Doc. 033.04.01, Dec. 2013.
[4] S. Sesia *et al.*, *LTE The UMTS Long Term Evolution: From Theory to Practice*, John Wiley & Sons, 2009.
[5] 3GPP. "Release 13 Work Item: IP Flow Mobility support for S2a and S2b Interfaces (NB-IFOM)," Available at: http://www.3gpp.org/DynaReport/WiSpec—640047.htm
[6] 3GPP TR 23.861 "Network based IP flow mobility V1.11.0," Nov. 2014.
[7] Open Networking Foundation, "SDN architecture v1.0." ONF technical reference document, June 2014.
[8] D. Farinacci *et al.*, "Locator/ID Separation Protocol (LISP)," IETF RFC 6830, Jan. 2013.
[9] OpenDaylight project; available at: http://www.opendaylight.org/
[10] K. Pentikousis, Y. Wang, and W. Hu. "Mobileflow: Toward software-defined mobile networks," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013.
[11] A. Basta *et al.*, "A Virtual SDN-Enabled LTE EPC Architecture: A Case Study for S-/P-Gateways Functions," *Proc. IEEE Software Defined Networks for Future Networks and Services*, Nov. 2013.
[12] C. Matsumoto, "Time for an SDN Sequel? Scott Shenker Preaches SDN Version 2," *SDNcentral*, Oct. 29, 2014.
[13] 3GPP TS 23.401, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access v13.0.0." Sept. 2014.
[14] A. Lara, A. Kolasani, and B. Ramamurthy. "Network Innovation using OpenFlow: A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, First Qtr. 2014.
[15] 3GPP TR 23.829, "Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO) v10.0.1," Oct. 2011.
[16] A. Krendzel, M. Portoles, and J. Mangues, "Fundamental Analysis of the EPS Architecture Facing Future Data Networking Needs," CTTC technical report, Nov. 2012; available at: http://networks.cttc.cat/wp-content/uploads/sites/2/2013/05/future_eps.pdf

Biographies

Marc Portoles Comeras received his degree in telecommunications engineering from the Technical University of Catalonia (UPC) and is currently working as a software engineer at Cisco Systems Inc. participating in the development of the LISP protocol architecture. Before joining Cisco

he was a research engineer at the Centre Tecnolgic de Telecomunicacions de Catalunya (CTTC) where he participated in multiple R&D projects funded by the EU and the Spanish Government, such as IST-WIP, BeFEMTO, and SYMBIOSIS. His current research interests are on SDN and network virtualization solutions.

JOSEP MANGUES-BAFALLUY (josep.mangues@cttc.cat) received his M.Sc. (1996) and Ph.D. (2003) in telecommunications from UPC, including an Erasmus grant from ENSEEIHT, Toulouse, France. He is senior researcher and head of the communication networks division (networks.cttc.cat) of CTTC and coordinated its IP Technologies Area from 2003 to 2012. He regularly participates in EU (FP6, FP7, H2020), industrial (e.g., Cisco, Orange, Aviat Networks), Spanish, and Catalan research projects, of which six were led by him. His research interests include software-defined networks (SDN) and network functions virtualization (NFV) applied to self-organized mobile networks.

ANDREY KRENDZEL is a senior researcher at Huawei Technologies Oy, Finland, 5G Networks, since October 2013. From 1995 to 2013 he was involved in research activity related to xG (x = 2, 3, 4) networks at CTTC, Tampere University of Technology (TUT), and the Leningrad R&D Telecommunication Centre (LONIIS). He received his Ph.D. degree from TUT in 2005 and his licentiate degree (Candidate of Science) from St. Petersburg University of Telecommunications (SUT) in 2001.

MANUEL REQUENA ESTESO (1974) received his M.Sc. in computer science from the Technical University of Valencia (1998). He formulated his Master's thesis at ENSEA, France. He worked as a software engineer developing telecommunication software solutions (Atos Origin Integration, 1998–2003). Currently, he is the coordinator of the EXTREME Testbed in the Mobile Networks Department of CTTC. Since 2003, he has participated in industrial and publically funded projects in wireless, mobile, and optical networking.

ALBERT CABELLOS received his B.Sc. (2001), M.Sc. (2005), and Ph.D. (2008) degrees in computer science engineering from UPC. In September 2005 he became an assistant professor of the Computer Architecture Department. He is an Editor of the *Elsevier Journal on Nano Computer Network* and founder of the ACM NANOCOM Conference, the IEEE MONACOM Workshop, and the N3Summit. He has also founded the LISPmob open-source initiative along with Cisco. He has been a visiting researcher at Cisco Systems and Agilent Technologies, and a visiting professor at the Royal Institute of Technology (KTH) and the Massachusetts Institute of Technology (MIT). He has participated in several national (Cicyt), EU (FP7), U.S. (NSF), and industrial projects (Samsung and Cisco). He has given more than 10 invited talks (MIT, Cisco, INTEL, MIET, Northeastern University, etc.) and co-authored more than 15 journal and 40 conference papers. His main research interests are future architectures for the Internet and nano-scale communications.

ACCEPTED FROM OPEN CALL

Coverage Enhancement Techniques for Machine-to-Machine Communications over LTE

Ghasem Naddafzadeh-Shirazi, Lutz Lampe, Gustav Vos, and Steve Bennett

ABSTRACT

The tremendous growth of machine-to-machine (M2M) applications has been a great attractor to cellular network operators to provide machine-type communication services. One of the important challenges for cellular systems supporting M2M terminals is coverage, because terminals can be located in spaces in buildings and structures suffering from significant penetration losses. Since these terminals are also often stationary, they are permanently without cellular coverage. To address this critical issue, the third generation partnership project (3GPP), and in particular its radio access network technical specification group, commenced work on coverage enhancement (CE) for long-term evolution (LTE) systems in June 2013. This article reviews the CE objectives defined for LTE machine-type communication and presents CE methods for LTE downlink and uplink channels discussed in this group. The presented methods achieve CE in a spectrally efficient manner and without notably affecting performance for legacy (non-M2M) devices.

INTRODUCTION

Recent advances in the provision of reliable and ubiquitous cellular communications have paved the way for a new era in wireless communications that will see massive communication between automated devices over wireless links. This concept, commonly referred to as machine-to-machine (M2M) communications, is an enabling technology for a large variety of application domains, including electricity management (smart grids), healthcare, transportation and logistics, and home and industry automation. As a result, M2M communications is one of the fastest-growing technologies in the field of telecommunications. The GSM Association forecasts that the number of M2M cellular connections will reach approximately a quarter of a billion in 2014. Furthermore, according to ABI research from January 2014, the (projected) number of annual shipments of wireless modules for cellular M2M communications will increase from approximately 40 million in 2011 to approximately 185 million in 2019. This demonstrates the strong incentive for cellular wireless technology providers to participate in this market. The latest cellular communication standards developed by the third generation partnership project (3GPP) are long-term evolution (LTE) and LTE-advanced (LTE-A). Many cellular network operators are migrating from GSM, UMTS/HSPA, and other legacy standards to LTE. LTE provides a flexible communication architecture designed to enable communication at a lower cost per bit and to accommodate the continuous growth in wireless cellular demand, both in the number of connections and in the required data rate [1]. There is an obvious advantage for operators if the expanding LTE infrastructure will also support M2M applications (and eventually the Internet of Things (IoT)).

Wireless devices for M2M communication generally serve applications whose quality-of-service requirements are different from those handled by conventional (human-operated) LTE user equipment (UE). For example, many M2M applications require transmitting only infrequent and short messages and are often more delay-tolerant compared to human-to-human (H2H) and human-to-machine (H2M) applications. In this context, requirements on peak data-rates can often be relaxed. We expect that the market for wireless modules supporting only these lower rates ($<$ 1Mb/s) will grow notably faster than the total cellular M2M market.

Against this background, the 3GPP standardization process has recognized the need for extending the LTE standard to better support M2M applications and to meet the specific requirements of machine-type communication (MTC) devices. In particular, 3GPP has started to add MTC-specific optimization into LTE-A starting from Release Ten (Rel-10) of the standard in 2010 [2]. Several new MTC-related work items have been studied for Rel-12 [3]. These include the introduction of a new UE category,

Ghasem Naddafzadeh-Shirazi and Lutz Lampe are with the University of British Columbia.

Gustav Vos and Steve Bennett are with Sierra Wireless Inc.

0163-6804/15/$25.00 © 2015 IEEE

Category		1	2	3	4	5	0
Peak rate Mb/s	DL	10	50	100	150	300	1
	UL	5	25	50	50	75	1
Capability for physical layer functionalities							
RF bandwidth		20 MHz					20 MHz
Modulation	DL	QPSK, 16QAM, 64 QAM					All
	UL	QPSK, 16QAM				QPSK, 16QAM, 64QAM	QPSK, 16QAM
Multi-antenna							
2 Rx diversity		Assumed in performance requirements					Not supported
2x2 MIMO		Not supported	Mandatory				Not supported
4x4 MIMO		Not supported				Mandatory	Not supported

Table 1. Comparison of some features for LTE UE categories 1-5 and LTE MTC UEs (category 0 (CAT0) devices). DL = downlink, UL = uplink.

the so-called category 0 (CAT0), equipped with a single antenna, half-duplexing frequency-division duplex (HD-FDD), and lower transport block sizes in order to reduce costs [4].

Table 1 shows some of the properties of the new CAT0 UE in comparison to legacy LTE UE categories 1 to 5, which are dominantly used in current LTE network deployments. We observe that CAT0 UEs have reduced peak-rates and diversity/MIMO capabilities for the benefit of low-cost design. The associated estimated cost savings are of the order of 50 percent compared to CAT1 devices [3]. Further cost savings in terms of reduced bandwidth, maximum transmit power, and support for downlink transmission modes are investigated in a recently opened 3GPP work item for Rel-13 [5].

For a number of M2M applications, such as remote operation of vending machines, remote metering, or remote maintenance and control, MTC UEs can be installed inside buildings or structures with large penetration losses [6]. Furthermore, since these UEs are not mobile, they have no possibility of improving link quality. Hence, permanent coverage holes can occur. This critical shortcoming can only be overcome by *coverage enhancement* methods. In response to this, a coverage enhancement work item for MTC UEs was approved in the 3GPP radio access network technical specification group (RAN TSG) in June 2013 [4]. Initially, the aim was to complete this work item for Rel-12, but due to time limitations it was postponed and reopened in September 2014 for inclusion in Rel-13.

In this article we review the coverage enhancement (CE) targets specified for MTC LTE and present CE techniques that can provide cellular connectivity in adverse propagation conditions and are considered for inclusion in the MTC standardization. For the former we first briefly describe the LTE resource structure and the coverage in its uplink (UL) and downlink (DL) channels. Then we choose three different LTE channels and explain the possible solutions for coverage enhancement. These include novel methods that can provide flexible CE under different network conditions. They do not require modifications of legacy LTE UEs and have little effect on their performance (e.g. by way of limitations to resource scheduling). They also attempt to retain overall cell spectral efficiency, which is important from a cost per bit perspective for mobile network operators. Focussing on CE for MTC LTE, this article is complementary to the discussion of CE for LTE-advanced presented recently in [7].

Overview of Coverage in Uplink and Downlink LTE Channels

In this section we first briefly review LTE physical uplink and downlink channels and present a summary of the current coverage in these channels. This sets the stage for the coverage enhancement methods presented thereafter.

LTE Channels

In LTE, data is mapped to orthogonal radio resources in the time-frequency plane. The atomic data unit, known as a resource element (RE), has the symbol duration of 66.7 microseconds, which corresponds to a subcarrier bandwidth of 15 kHz. For a normal cyclic prefix length, a grid of 7×12 REs in the time-frequency domain is known as an LTE physical resource block (PRB). A PRB pair forms the basic unit commonly used in the LTE standard for scheduling and resource allocation. Taking guard bands and cyclic prefix into account, a PRB occupies approximately 200 kHz over half a

A PRB pair forms the basic unit commonly used in the LTE standard for scheduling and resource allocation. Taking guard bands and cyclic prefix into account, a PRB occupies approximately 200 kHz over half a millisecond, which is also the duration of an LTE time slot.

Maximum coupling loss is a measure of coverage in LTE channels. It is defined as the difference between maximum transmission power in the channel and its corresponding receiver sensitivity. A higher MCL value indicates a smaller required signal-to-noise ratio for a target (often one percent) block error rate, which translates into better coverage for that channel.

MCL in dB	UL channels			DL channels		
	PUCCH	PRACH	PUSCH	PDCCH	PBCH	PDSCH
Category 1 UE	147.2	141.7	140.7	146.1	149.0	145.4
Category 0 UE	147.2	141.7	140.7	142.1	145.0	141.4
Target MCL	155.7	155.7	155.7	155.7	155.7	155.7
Required CE for CAT0 UE in dB	8.5	14.0	15.0	13.6	10.7	14.3

Table 2. MCL for UL and DL LTE channels in FDD mode [3, 8]. eNodeB in two transmit and two receive antenna configuration. UE CAT1 with one transmit and two receive antennas, UE CAT0 with one transmit and one receive antenna.

millisecond, which is also the duration of an LTE time slot. One LTE sub-frame consists of two time slots (1 ms), and 10 consecutive sub-frames form a radio frame.

In both the UL and DL directions, there are different physical channels, which are transmitted in specific REs of the time and frequency radio resources. The physical DL and UL shared channels (PDSCH and PUSCH) are dedicated to data exchange between the LTE base station (eNodeB) and the UEs. The size of the medium access control protocol data unit is called the transport block size (TBS), and the time taken for its transmission is referred to as the data transmission time interval (TTI). The TTI is equal to the duration of one sub-frame. The data channels are complemented by a number of control channels, including the physical DL control channel (PDCCH) for allocating PRBs to PDSCH and PUSCH, the physical UL control channel (PUCCH) for transmitting UE resource requests and link quality information, the physical broadcast channel (PBCH) in the DL, which broadcasts the information required at a UE for joining a cell, and the UL physical random access channel (PRACH), which is used for contention-based random access for requesting a resource allocation from the eNodeB.

Coverage Requirements

Maximum coupling loss (MCL) is a measure of coverage in LTE channels. It is defined as the difference between maximum transmission power in the channel and its corresponding receiver sensitivity [8]. A higher MCL value indicates a smaller required signal-to-noise ratio (SNR) for a target (often one percent) block error rate (BLER), which translates into better coverage for that channel.

The 3GPP study item [8] focused on identifying the LTE channels with critical MCLs. For this, the study item considered medium data rata and VoIP applications. Table 2 summarizes the MCL of the above-mentioned channels in LTE as reported in [3] and [8]. Since CAT0 UEs will be equipped with only one receive antenna as shown in Table 1, a 4 dB penalty has been applied to the MCL of downlink channels in Table 2. Furthermore, a target MCL of 155.7 dB has recently been agreed on for CAT0 UEs [9]. The resulting required CEs for the different channels are summarized in the last row in Table 2.

Solutions suggested to achieve CE include signal repetition and/or more efficient detection and decoding techniques, relaxed reception requirements, new channel and signal design, and power boosting [4]. In the remainder of this article we elaborate on such methods. Furthermore, since data usage for MTC UEs is far lower than for typical H2H devices, spectral efficiency for UE-specific traffic is less of a concern compared to broadcast information, such as master information blocks (MIBs) and system information blocks (SIBs), which need to be continuously sent in LTE systems. Given this, we focus on CE techniques to improve the PBCH, which is dedicated to MIB transmission, and SIB broadcasting, which is scheduled by the PDCCH and sent via the PDSCH. Another characteristic of MTC UEs is that they tend to send UL data much more often than DL data. Therefore, and since the PUSCH requires the largest coverage gain (see Table 2), we also focus on novel CE techniques to improve the PUSCH.

We start with the PBCH in the next section and demonstrate the possibility of a 10.7 dB CE through novel decoder designs. After that we present CE for SIB broadcasting. Finally, a transmission strategy based on spreading and bundling of data is introduced in order to efficiently achieve a 15 dB CE in the PUSCH. These strategies generally exploit the relaxed MTC latency requirements by prolonging the decoding time and having the MTC UE waiting for more data. If stricter latency requirements apply, CE often leads to a reduced spectral efficiency. We note that the methods presented in this manuscript have been introduced to the RAN TSG by the authors in [10–12]. At the time of this writing, the RAN TSG is considering these MTC CE techniques for possible inclusion in the Rel-13 of the LTE standard. They are collated here with the aim of providing timely information to researchers and scholars interested in LTE MTC coverage enhancement.

Coverage Enhancement Techniques for PBCH

Since re-designing broadcasted channels, such as the PBCH, would break backward compatibility with legacy UEs, broadcasted channels need to be supported by a network for a long time. At

the same time, decoding the PBCH is the prerequisite of a successful connection in low coverage. Hence, although the MCL value of the PBCH is better than those for other channels in Table 2, a 10.7 dB CE still needs to be achieved as efficiently as possible.

PBCH Background

The PBCH is transmitted in the first sub-frame of each frame and has a TTI of 40 ms. The PBCH nominally transmits 14 bits of control information (the MIB). Via cyclic-redundancy check (CRC) and a tail-biting rate-1/3 encoding and cyclic rate matching, codewords of 1920 bits are generated. These are divided into four redundancy versions (RVs), which are transmitted every 10 ms in sub-frame #0 of a radio frame. The relative location of the RV within the 40 ms TTI encodes another two bits of data (see [1] for details).

Due to the mentioned legacy issues, changes in the structure of the PBCH for CE would be impractical. Therefore, in the following we present two UE-based methods that achieve CE exploiting the existing PBCH structure [10].

Increase PBCH Decoding Attempts (IPDA) Method

The first CE method, which we refer to as the IPDA method, is to continue decoding PBCH transmissions until decoding has been successful. This conceptually simple method does not require any modifications to the PBCH, but rather relaxes the BLER target and the UE acquisition time. Furthermore, and different from other CE approaches such as PBCH repetition or power boosting [13], no additional spectrum or power resources are exhausted. However, IPDA can readily be combined with these methods. Finally, the IPDA method does not increase signal buffering or processing requirements of the UE. However, it significantly increases decoding latency, but this is often acceptable for MTC UEs and will be discussed later in the evaluation section.

Correlation Decoder (CD)

Our second method replaces the default PBCH decoder, consisting of de-rate-matching, tail-biting, and CRC decoding (see [1]), by a maximum-likelihood (ML) decoder. ML decoding is achieved by correlating the PBCH received samples with all possible PBCH sequences. Hence, we refer to the method as correlation decoder (CD). The number of possible sequences is $3 \times 4 \times 12288$, where the factors 3, 4 and 12288 are due to the possible three antenna configurations at the eNodeB, the four relative locations of the RV (2-bit information), and the fact that only 12288 of the 214 14-bit patterns are possible. The PBCH sequence with the highest correlation value is the ML estimate. However, to limit the possibility of false alarm, usually done through CRC decoding, the ratio of the likelihoods for the ML and the second-best sequence is compared to a threshold. If the ratio is above the threshold, the ML estimate is used as the decoding result, otherwise a decoding failure is declared.

The CD approach can naturally be extended to cover multiple TTIs. Furthermore, decoding can be performed after de-rate matching, when the PBCH sequence length becomes 120 samples for each TTI. This simplifies correlation to 120 additions per PBCH sequence.

Evaluation

Table 3 shows the CE simulation results for the IPDA and CD methods for an extended pedestrian type-A (EPA) channel with 1 Hz Doppler. For the simulations, we used the MATLAB LTE toolbox and the settings listed in [10]. For the IPDA case, we also consider its combination with intermittent duplication of the PBCH (as suggested in [14]) in order to reduce acquisition time. The table shows the acquisition times for different percentiles (99 percent, 90 percent, and mean) and for different coverage gains, which are calculated based on a required one percent BLER for PBCH.

We observe that IPDA can provide enhanced coverage at the expense of acquisition time. Acquisition time can be reduced through intermittent duplication, which requires additional resources and thus increases unwanted overhead. However, the PBCH acquisition times drop sharply when less than the maximal CE of 10.7 dB is required. This means that only UEs in the deepest coverage holes will experience occasional lengthy acquisition times.

For the CD the table shows the CE for different correlation lengths and thus acquisition times (which are multiples of the 40 ms TTI). For an acquisition time of 160 ms and one percent BLER (i.e. 99 percent success rate), the CD method can provide an approximately 9.5 dB additional gain compared to the IPDA method without repetition, because the CD is the optimal ML decoder. While the CD is more complex than IPDA, it can be made simpler by noting that during re-acquisition some of the MIB values are known, which reduces the number of required correlations. Thus, MTC UEs may implement both the IPDA method and the CD, the former for initial MIB acquisition and the latter for re-acquisition of MIBs.

SIB Coverage Enhancement Using Restrictive SIB Scheduling Methods

Like the MIB sent on the PBCH, SIBs are also broadcast by the network, and extra care must be taken in the broadcast design, because there will be no opportunity to improve it in the future without breaking backward compatibility. Although simple repetition for the PDCCH and PDSCH can be used to provide the required CE for SIBs, this would result in a spectrally inefficient implementation. In this section we describe alternative methods, which we collectively refer to as restrictive SIB scheduling and which provide coverage gain in a spectrally efficient manner [11].

SIB Background

In the current Rel-12 there are 19 different SIBs that are broadcast. However, since CE is desired for mostly stationary UEs, many of the 19 SIBs

Like the MIB sent on the PBCH, SIBs are also broadcast by the network, and extra care must be taken in the broadcast design, because there will be no opportunity to improve it in the future without breaking backward compatibility.

	Duplication intermittency	Coverage gain	Mean acquisition time (ms)	90th percentile acquisition time (ms)	99th percentile acquisition time (ms)
IPDA method	Repetition sent every frame (100 percent more PBCH resources)	10.7 dB	65.1	120	600
		6.0 dB	43.5	80	240
		3.0 dB	41.0	80	120
	Repetition sent every 2nd frame (50 percent more PBCH resources)	10.7 dB	77.3	120	720
		6.0 dB	46.0	80	280
		3.0 dB	41.7	80	120
	Repetition sent every 4th frame (25 percent more PBCH resources)	10.7 dB	90.3	200	800
		6.0 dB	47.1	80	320
		3.0 dB	41.8	80	160
	No repetition	10.7 dB	105.8	240	1000
		6.0 dB	49.6	120	400
		3.0 dB	43.2	80	160
Correlation decoder	No repetition (1 percent BLER)	2.3 dB	40 (1 TTI)		
		4.5 dB	80 (2 TTIs)		
		7.5 dB	120 (3 TTIs)		
		12.5 dB	160 (4 TTIs)		

Table 3. CE achieved with IPDA and CD as a function of the acquisition time. CD uses a fixed decoding window of n TTIs ($n = 1, 2, 3, 4$), and thus acquisition time is a constant equal to $n \times 40$ ms.

do not need to be decoded. In fact, only SIB1, SIB2, and SIB14 need to be decoded by the UE when utilizing CE. If other SIBs are required, they could be sent via unicast methods. As SIB1 is the most important SIB, it is sent most frequently (every 20ms) and at a known sub-frame (SF) (SF#5 of every other frame), and it must be decoded by the UE before the other SIBs.

Unlike MIBs, which have a dedicated physical channel (i.e. the PBCH), SIBs use the same physical channels as user plane data, that is, the PDCCH for scheduling and the PDSCH for the data. We will discuss methods for CE for both channels in the following.

Combining-Legacy-SIBs-PDSCH Method

In the current LTE standard, different levels of coverage for the SIBs can be obtained by changing a SIB's coding rate or the number of repeats that are sent. Given there is a limited number of PRBs that can be used in an SF (e.g. only six in a 1.4 MHz system), repeating SIBs is an important method used today to extend coverage. We note that all the repeats should be sent within the so called system-information (SI) window. However, the information in SIB1, SIB2, and SIB14 is often static for long periods of time in normal operating conditions (e.g. the network is not in an emergency overload situation). Hence, to enhance coverage for the data portion of SIBs, that is, the PDSCH, the UE could combine SIB repetitions beyond the SI window. Results in [11] show that for a 15 dB CE for SIB1, the 99th percentile of the acquisition time is 2.4 seconds (corresponding to combining 120 copies of SIB1). A similar number of copies would be expected for other SIBs. Since the other SIBs of interest (SIB2 and SIB14) are not sent as often, the acquisition time for those SIBs would typically be longer. In the context of MTC, we consider an extended acquisition time as generally more acceptable compared to increasing SIB transmission periodicity, causing a loss in spectral efficiency. We also note that since the SIB message contains a CRC, the UE will stop decoding when it has correctly decoded the SIB. Thus, for the example above, acquisition time will often (i.e. 99 percent of the time) be shorter than 2.4 seconds.

PDCCH-Less SIB Decoding Method

As mentioned above, the PDCCH is required to be decoded to obtain scheduling information. However, lowering code rates for each PDCCH message is not sufficient to provide an up to 15 dB CE, because there are not enough PDCCH resources in an SF. Thus, the PDCCH message would have to be repeated across many SFs,

which is spectrally inefficient. To avoid this, we first explore whether it would be possible to decode the mentioned SIBs without prior PDCCH decoding. Such a PDCCH-less SIB decoding method would then require a different mechanism for the UE to acquire all the information contained within the PDCCH, so that the UE can skip PDCCH decoding. In particular, the UE needs (1) SIB transmission timing, (2) SIB PRB locations within the band, and (3) SIB coding rate.

SIB Transmission Timing: For SIB1, there is no problem because the SIB1 transmission timing is already known (e.g. every other SF#5). Thus, SIB1 could be used to also provide the precise transmission timing (i.e. more than the SI window already transmitted via SIB1) for SIB2 and SIB14. This would restrict the eNodeB's scheduler in that it may have to postpone a UE's DL transmission and the UE may experience additional latency. However, there is no decrease in spectral efficiency due to the scheduling restrictions.

SIB PRB Locations within the Band and Coding Rate: An intriguing solution is to provide the PRB location and coding rate within the 10 available spare bits of the MIB (legacy UEs will ignore these spare bits). Since the MIB has limited capacity, only the SIB1 information needs to be in the MIB, and similar to above, SIB1 can carry this information for SIB2 and SIB14 (legacy UEs will ignore these new information elements in the SIB1). Although this method still allows the PRB location and coding rate to be dynamic, the eNodeB loses some scheduling flexibility.

Combing-PDCCH Method

An alternate method to PDCCH-less SIB decoding is to have the UE combine the copies of the PDCCH message already being sent to legacy UEs. For this to be feasible, the content of the PDCCH message must be static and the location of the PDCCH message within the PDCCH must be known a-priori to the UE. For the content of the PDCCH message to be static, the PRB location and coding rate of the SIB must be static.

This method requires the UE to complete two steps: decode the PDCCH and then decode the PDSCH. Thus this method would take longer to achieve SIB reception than the PDCCH-less method given the same number of repetitions. Furthermore, if additional repetitions are sent to reduce the SIB acquisition time, the PDCCH and PDSCH portions would need to be repeated. Like the PDCCH-less solution, this solution will be backward compatible since fixing the PDCCH and its contents will be transparent to legacy UEs.

TTI Bundling and CDMA for Coverage Enhancement in the PUSCH

Many MTC UEs will have dominantly UL transmission. Furthermore, as shown in Table 2, the UL data channel, that is, the PUSCH, requires the largest CE to meet the coverage target. Therefore, our final CE considerations concern the PUSCH.

PUSCH Background

Data transmitted over the PUSCH is encoded with a rate-1/3 turbo encoder and then rate-matched and arranged in four RVs, each of which matches the TBS (see [1]). Based on this, incremental-redundancy automatic-repeat request (ARQ) can be performed after each TTI. That is, the receiver acknowledges the receipt of data, and in the case of a negative Acknowledgment, the next RV of the current data will be transmitted. The default schedule for PUSCH transmission is to transmit one RV in one TTI, and to only transmit another RV if requested via negative acknowledgment (NACK). The authors in [8] tackle the issue of CE for the PUSCH through TTI bundling. In TTI bundling, all RVs are transmitted at once, without waiting for a NACK. This leads to CE for delay-limited application such as VoIP. The current LTE standard assumes a fixed bundling size of four, but the current work in 3GPP for VoIP and medium data rates considers increasing the TTI bundle size to higher values to provide modest coverage gains.

Since M2M applications are often delay-tolerant, we can think of combining bundling with repetition for CE. For example, increasing the bundling size from four to eight means the UE would send each RV twice [15]. Furthermore, the bundling size could be adjusted dynamically, considering the UE's need for CE and its delay tolerance [15].

An alternative to repetition of data is the use of spreading. The advantage of spreading is that it enables multiple UEs to transmit concurrently, that is, to perform code-division multiple-access (CDMA), which in turn improves system spectral efficiency. CDMA is already used in LTE, namely for PUCCH format-3 to provide multiple-user access on the control channel [1].

In the following, we explain a method for simultaneous use of adaptive TTI bundling and spreading for MTC CE, and we present a signaling procedure for the flexible assignment of PUSCH resources to a variable number of UEs [12].

Flexible TTI Bundling with CDMA Support

Our method extends conventional LTE TTI bundling by adjusting the bundling size and the spreading factor used by UEs, according to the instantaneous cellular network conditions. These are defined through the number of "active" MTC UEs, that is, UEs that have data to transmit, their channel quality and thus instantaneous coverage, and the available PUSCH resources. The main advantage of using flexible bundling and spreading is that CE is achieved without overly compromising network spectral efficiency. Spreading is performed over REs at the same frequency, which simplifies despreading assuming the channel remains essentially constant over the spreading interval. Denoting the spreading length by N_S and performing bundling with bundling size N_B, the CE offered by the flexible TTI bundling and CDMA is about $10\log10(N_B \times N_S)$ dB. The exact gain is somewhat larger when combining different RVs from

An alternative to repetition of data is the use of spreading. The advantage of spreading is that it enables multiple UEs to transmit concurrently, that is, to perform CDMA, which in turn improves system spectral efficiency. CDMA is already used in LTE, namely for PUCCH format-3 to provide multiple-user access on the control channel.

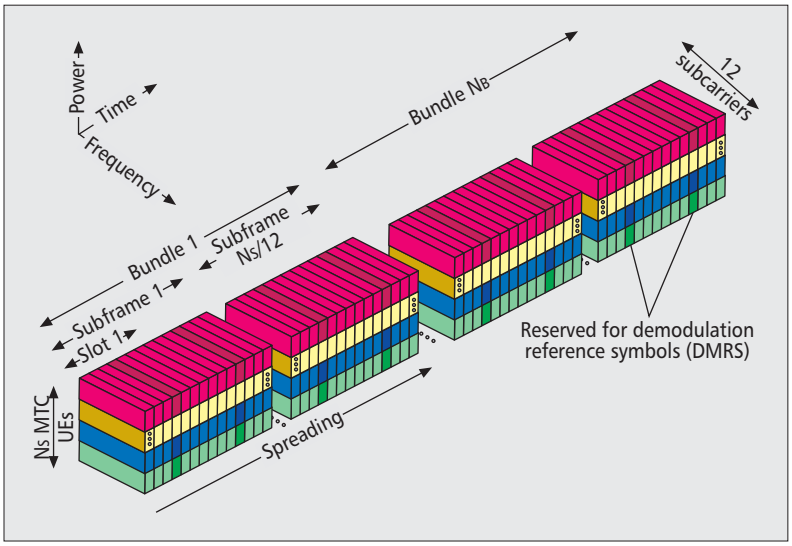


Figure 1. Structure of the bundling and spreading blocks for flexible TTI bundling and CDMA in PUSCH. Spreading with length N_S is performed over one or more consecutive sub-frames, which form a "spreading bundle." The spreading bundle is repeated N_B times. Calculation and assignment of N_B, N_S, and code index to MTC UEs is done by a scheduling protocol in the eNodeB.

the turbo code contained in TTI bundles. Figure 1 shows the structure of the code-spread TTI bundling for the PUSCH.

Protocol for Flexible TTI Bundling and CDMA

In order to successfully schedule the active MTC UEs to transmit in this scheme, the eNodeB should first adjust the values of N_S and N_B based on the number of active UEs and the required coverage gain, respectively. Then it informs the MTC UE of the values of N_B and N_S and the assigned codes. To minimize the impact of this procedure on the current LTE standard, we note that some of the existing control flags in the PDCCH uplink grant are unlikely to be used in the MTC mode. Thus, they can be reused to inform the UE to obtain its TBS, bundling size, spreading length, and code index. This would be done based on a configuration table for flexible TTI bundling and CDMA, which is a modified version of the TBS table used in legacy UEs. Using a modified TBS table, transmission with flexible TTI bundling and CDMA can be scheduled as follows:

1. When data is available for transmission, the MTC UE sends a scheduling request on the PUCCH.
2. The eNodeB waits for a predefined time, collecting requests of MTC UEs as in Step 1, and estimates their required coverage gain from the received channel quality index (CQI).
3. The eNodeB sets N_S to the closest spreading length that is available in the configuration table such that the current number of active MTC UEs can be accommodated.
4. The eNodeB chooses N_B based on the required coverage gain, $CE=10\log_{10}(N_B \times N_S)$ with N_S from Step 3.
5. Based on Steps 3 and 4 and available resources, the eNodeB assigns resources to UEs. It sends a PDCCH DCI format-0 for PUSCH allocation and sets a flag to indicate that the modified TBS table needs to be used.

Waiting and collecting requests in Step 2 is done to utilize as much of the available CDMA code as possible, which maximizes system spectral efficiency while providing CE through spreading for individual MTC UEs. Steps 3 and 4 can further be refined to account for MTC UEs in good coverage, which may not need the full spreading gain. For example, those UEs could be assigned shorter spreading sequences or multiple longer spreading sequences.

Performance Evaluation

According to the LTE coverage enhancement study [8], we evaluate the CE achieved with flexible bundling and CDMA by measuring the SNR required for a BLER of two percent. The first three columns of Table 4 show selected (N_S,N_B) combinations and the expected coverage gains compared to the case without bundling and spreading. We observe the CE due to spreading and bundling, which is proportional to $N_B \times N_S$, reaches the required 15 dB (Table 1) for various parameter combinations. Column 4 shows simulated coverage gains for the EPA channel with 1 Hz Doppler, assuming a TBS of 104 bits transmitted in one PRB (i.e. 180 kHz bandwidth) using QPSK and CDMA with orthogonal spreading sequences. The theoretical and simulated gains match well, where the latter include the effect of combining RVs in the Turbo decoder, which gives only another approximately 0.4 dB gain compared to pure repetition due to the already low code rate for only one RV. The last two columns of Table 4 show the spectral efficiency (over all MTC UEs) and data rate (per MTC UE). As can be seen, system spectral efficiency is affected by bundling but not by spreading, assuming that all spreading codes are used. Hence, no resources are wasted while benefitting from spreading gain and achieving the required CE. However, the amount of spreading that can be applied is limited by the need for an essentially time-invariant channel over N_S PRBs or extra reference symbols for channel estimation.

Conclusion

This article has reviewed recent efforts presented in 3GPP to enhance coverage for LTE MTC. We have described methods for downlink broadcast and uplink data channels that are able to meet the CE targets specified for CAT0 devices. An overview of the presented CE methods and their effects on the LTE system is provided in Table 5. They build on the existing LTE signal structures and are thus backward compatible, affect legacy and non-MTC UEs as little as possible, and maintain high system spectral efficiency. The latter is particularly relevant for broadcast channels. The achieved CE generally comes at the cost of increased latency of transmission, which is a natural trade-off for enhanced coverage and acceptable for many MTC applications. We believe that amendments to the LTE standard that support

Spreading length N_S	# of TTIs bundled N_B	Theoretical CE (dB) $10\log_{10}(N_B \times N_S)$	Simulated CE (dB) (perfect channel estimation)	Spectral efficiency over all MTC UEs bps/Hz	Data rate per MTC UE (kbs)
1	1	0.0	0.0	0.578	104.0
2	6	10.8	11.2	0.097	8.7
6	2	10.8	10.7	0.290	8.7
22	2	16.4	16.4	0.289	2.4
12	9	20.3	20.6	0.067	1.0
1	66	18.2	18.6	0.009	1.6
66	1	18.2	18.1	0.578	1.6
1	72	18.6	19.0	0.008	1.4

Table 4. Achieved coverage gain and spectral efficiency for flexible TTI bundling and spreading. Simulated CE, spectral efficiency and data rate for TBS = 104 in one PRB using QPSK.

Channel	CE method	Pros (+) and cons (–) to achieve CE	
PBCH	IPDA	+ No changes to legacy broadcast channel + No changes to spectrum and power efficiency of system	+ No extra buffering or processing at UE – Longer acquisition time
	CD		– Increased processing at UE
PDSCH for SIB	SIB-PDSCH combining	+ Fully compatible with legacy UEs + No changes to spectrum and power efficiency of system	– Longer acquisition time – Need static SIB messages
PDCCH for SIB	PDCCH-less SIB decoding		– eNodeB scheduling restrictions – Use spare bits of MIB
	PDCCH Combining		– Longer acquisition time – Need known PDCCH location and static SIB location and code rate
PUSCH	TTI bundling with CDMA	+ Better spectrum and power efficiency than repetition – Loss in system spectrum efficiency if few UEs in poor coverage – Boosting or repetition of reference symbols may be required	

Table 5. Summary of presented CE methods and their effects on the LTE system.

CE for low-cost CAT0 UEs are very important to ensure that LTE will be competitive with alternative wireless access technologies, such as custom IoT wide-area network protocols (e.g. Weightless™ and SigFox™), which claim to offer very low-cost devices, high coverage, and very good battery life. Further M2M and IoT changes for LTE (e.g. to improve LTE battery life) are being discussed in 3GPP, and we expect to see them in Releases 13 and beyond.

Acknowledgment

The authors would like to thank MITACS Canada for supporting this work through a MITACS Elevate post-doctoral fellowship.

References

[1] M. Baker, S. Sesia, and I. Toufik, *LTE — The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed.: Wiley, 2011.
[2] Third Generation Partnership Program (3GPP) Technical Specification 22.368, V.13.0.0, "Service Requirements for Machine-Type Communications (MTC); Stage 1," Jun. 2014.
[3] Third Generation Partnership Program (3GPP) Technical Report 36.888, V.12.0.0, "Study on Provision of Low-Cost Machine-Type Communications (MTC) User Equipments (UEs) Based on LTE," Jun. 2013.
[4] Third Generation Partnership Program (3GPP) Work Item Description RP-140522, "Low Cost & Enhanced Coverage MTC UE for LTE," http://www.3gpp.org/DynaReport/FeatureOrStudyItemFile-600012.htm, Jun. 2013.
[5] Third Generation Partnership Program (3GPP) Work Item Description RP-141660, "Further LTE Physical Layer Enhancements for MTC," Sep. 2014.
[6] C. Hägerling, C. Ide, and C. Wietfeld, "Coverage and Capacity Analysis of Wireless M2M Technologies," *Proc. IEEE Int'l Conf. Smart Grid Communications* (SmartGridComm), 2014, pp. 374–79.
[7] Y. Yuan *et al.*, "LTE-Advanced Coverage Enhancements," *IEEE Commun. Mag.*, vol. 52, no. 10, Oct. 2014, pp. 153–59.
[8] Third Generation Partnership Program (3GPP) Technical Report 36.824, "Evolved Universal Terrestrial Radio

Access (E-UTRA); LTE Coverage Enhancements, V.11.0.0," Jun. 2012.
[9] Ericsson, Alcatel-Lucent, Alcatel-Lucent Shanghai Bell, AT&T, DTAG, Intel, InterDigital, KDDI, KT, LG Electronics, Nokia Networks, Nokia Corporation, Panasonic, Qualcomm, Sierra Wireless, Sony, 3GPP Technical Document R1-145384, "WF on Coverage Enhancement Targets for MTC," Nov. 2014.
[10] Sierra Wireless, 3GPP Technical Document R1-144601, "Coverage Enhancement PBCH Simulation Results and Proposals," Nov. 2014.
[11] Sierra Wireless, 3GPP Technical Document R2-140215, "Study on Combining Legacy SIBs for MTC Coverage Enhancement," Feb. 2014.
[12] Sierra Wireless, 3GPP Technical Document R1-125082, "MTC Coverage Improvement through Variable TTI Bundling and Variable Length Code Spreading," Nov. 2012.
[13] Samsung, 3GPP Technical Document R1-131015, "PBCH Coverage Enhancements for Low-Cost MTC UEs," Apr. 2013.
[14] Alcatel-Lucent 3GPP Technical Document R1-130938, "PBCH Coverage Extension for MTC Devices," Apr. 2013.
[15] Huawei, HiSilicon, 3GPP Technical Document R1-121005, "Further Discussion on Coverage Enhancement," Mar. 2012.

Biographies

Ghasem Naddafzadeh-Shirazi received his B.Sc. degree in computer science and engineering (CSE) as the top-ranking student from Shiraz University, Iran, in 2006, his Masters degree in electrical and computer engineering (ECE) from the National University of Singapore in 2009, and his Ph.D. degree in ECE from the University of British Columbia (UBC), Canada, in 2014. He is currently a post-doctoral fellow at UBC and Sierra Wireless. His research interests include optimization for wireless communication networks, including long-term evolution (LTE) and ultra-wideband (UWB), and the application of machine learning in wireless sensor networks.

Lutz Lampe [M'02, SM'08] received Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from the University of Erlangen, Erlangen, Germany, in 1998 and 2002, respectively. Since 2003 he has been with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada, where he is a full professor. His research interests are broadly in theory and application of wireless, optical wireless, and power line communications.

Gus Vos is a chief engineer for Sierra Wireless. He serves as the company's 3GPP representative and is actively involved in developing and directing the required changes to the LTE standard to optimize it for the Internet of Things. He has 25 years of experience in wide-area wireless communications, having started his career working with proprietary protocols before moving to his current role in advocating for standardized cellular protocols. He holds a B.Eng. from the University of Victoria, Canada, and an M.Eng. from Simon Fraser University, Canada.

Steven Bennett (sbennett@sierrawireless.com) received his B.Sc.(Eng) degree from the University of London, U.K. in 1981. Since 2008 he has been at Sierra Wireless working on new technology in the CTO department. He participates in the 3GPP standardization body (www.3gpp.org), LTE working group RAN2. He has varied experience in R&D design for wireless communications products at the system and circuit levels. He has designed RF, analog, and digital circuits, integrated circuits and coded for DSP and embedded control. He has 22 patents.

ACCEPTED FROM OPEN CALL

LISP: A Southbound SDN Protocol?

Alberto Rodriguez-Natal, Marc Portoles-Comeras, Vina Ermagan, Darrel Lewis, Dino Farinacci, Fabio Maino, Albert Cabellos-Aparicio

ABSTRACT

The Locator/ID Separation Protocol (LISP) splits current IP addresses overlapping semantics of identity and location into two separate namespaces. Since its inception the protocol has gained considerable attention from both industry and academia, motivating several new use cases to be proposed. Despite its inherent control-data decoupling and the abstraction and flexibility it introduces into the network, little has been said about the role of LISP on the SDN paradigm. In this article we try to fill that gap and analyze if LISP can be used for SDN. The article presents a systematic analysis of the relevant SDN requirements and how such requirements can be fulfilled by the LISP architecture and components. This results in a set of benefits (e.g. incremental deployment, scalability, flexibility, interoperability, and inter-domain support) and drawbacks (e.g. extra headers and some initial delay) of using LISP for SDN. In order to validate the analysis, we have built and tested a prototype using the LISPmob open-source implementation.

INTRODUCTION

The Locator/ID Separation Protocol (LISP) [1] decouples identity from location on current IP addresses by creating two separate namespaces: endpoint identifiers to identify hosts, and routing locators to route packets. The original purpose of LISP was to solve the scalability issues of the Internet default-free zone (DFZ) routing tables by pushing traffic engineering practices to the identifiers space while keeping the locators space quasi-static and highly aggregatable. At the time of this writing LISP has been deployed in a pilot network (lisp4.net) that includes more than 20 countries and hundreds of institutions. LISP hardware and software are also widely available, both in open-source (lispmob.org, openlisp.org) and proprietary implementations (lisp.cisco.com).

Since its inception, LISP has gained significant traction in both industry and academia. As a result of LISP standardization and research efforts, the protocol has grown architecturally and has been applied to use cases beyond its original purpose. There is a growing interest in the role of LISP in Software Defined Networking (SDN) [2]. LISP is already becoming part of SDN solutions, such as the OpenDaylight controller (opendaylight.org). In this article we analyze the relation between the LISP architecture and the SDN paradigm.

There are two well-defined parts in any SDN deployment: the northbound and the southbound interfaces. The northbound offers a high-level application programming interface, where control applications can be deployed. The southbound is a low-level interface used to operate with the raw network elements. Currently, there is ongoing effort to define the high level abstraction interface (see Frenetic [3] or Procera [4] as examples). There are also several options with respect to the southbound interface, with OpenFlow [5] attracting the most interest from industry.

The main contribution of this article is to analyze LISP as a southbound SDN protocol. For this, the article presents a systematic analysis of the fundamental SDN requirements, inferred from the literature [2–10], and how such requirements can be fulfilled by the LISP architecture and components. The analysis results in a set of qualitative advantages and drawbacks as well as recommended potential improvements to overcome the identified issues. In order to validate the analysis, we build and test a prototype using the LISPmob open-source implementation (lispmob.org).

BACKGROUND: LISP OVERVIEW

The Locator/ID Separation Protocol (LISP) decouples host identity from its location. It creates two different namespaces: endpoint identifiers (EIDs) and routing locators (RLOCs). Hosts are identified by an EID, and their point of attachment to the network by an RLOC. To keep LISP incrementally deployable, in its very basic form EIDs and RLOCs are syntactically identical to current IPv4 and IPv6 addresses. However, the protocol allows arbitrary address families (e.g. MAC) to be used.

Figure 1 depicts the LISP common operation. Packets are routed based on EIDs within host sites and on RLOCs on transit networks. Since host A and host B are in different sites (e.g. two offices geographically separated), the packets from A to B have to traverse a transit network

Alberto Rodriguez-Natal and Albert Cabellos-Aparicio are with Technical University of Catalonia.

Marc Portoles-Comeras, Vina Ermagan, Darrel Lewis, and Fabio Maino are with Cisco Systems, San Jose, CA.

Dino Farinacci is with lispers.net.

0163-6804/15/$25.00 © 2015 IEEE

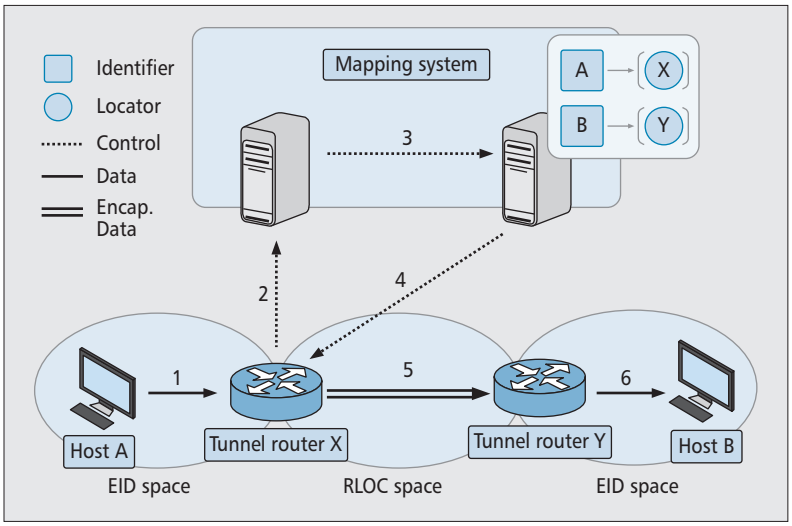


Figure 1. LISP overview.

(e.g. the Internet). To allow transit between the EID and RLOC spaces, LISP follows a map-and-encap approach performed by LISP tunnel routers deployed at edge points. In the image, tunnel router X receives the packet from host A addressed to host B (1). It knows that host B is in a different EID site, but it does not know where to reach that site (i.e. its RLOC). Tunnel router X requests this information to the mapping system (2). The LISP mapping system is a distributed database that stores EID to RLOC mappings. Tunnel router Y has previously registered its location and the set of EIDs it is in charge of in one of the mapping system internal servers. The mapping system routes the request internally (3) to find that server, and eventually it replies back with the requested location (4). Tunnel router X gets this information and caches it for future use. From now on, all EID packets from host A to host B will be encapsulated into an RLOC packet in tunnel router X and routed toward tunnel router Y (5). Upon arrival at the destination, tunnel router Y will decapsulate the packets and forward them natively to host B (6).

LISP: An SDN Architecture?

In this section we analyze if the LISP architecture, in its current form, can fulfill the requirements stemming from the SDN paradigm. Even though a formal definition of such requirements cannot be found in the literature, we infer the key SDN requirements by revisiting the design principles of the state-of-the-art SDN literature.

Control-Data Decoupling: One of the main reasons that motivated the emergence of OpenFlow [5] was to decouple the network control from the data forwarding devices. With its mapping system in place, LISP is capable of maintaining a distributed database where the network state and control information are stored. This database can be updated and queried by the LISP network elements in real time, and any change on it is propagated over the network. With this approach LISP is effectively decoupling control from data: while the data-plane remains at the router level, implemented on the tunnel routers, all control is pushed to the mapping system.

Network Programmability: Frenetic [3] and Procera [4] are two examples of the interest of the community in programing the network and improving its management. The LISP paradigm does not program the network but rather the mapping system. The control policies can be programmed and stored on the mapping system, then the LISP data-plane will operate accordingly. LISP semantics are poor when compared to state-of-the-art languages [3, 4] and focuses on representing the network state, therefore LISP should be complemented by a rich northbound language.

Centralized Control: Levin *et al.* [7] demonstrate that one core benefit of SDN is that it enables the network control logic to be designed and operated on a global network view, as though it were a centralized application. Since the LISP mapping system stores all network control state data and can be remotely accessed and updated in real time, it provides a global view of the network that effectively centralizes the control.

Scalability: Yeganeh *et al.* [8] show the concern of the SDN community about SDN scalability. LISP is a pull-based architecture that stores the network state information in the mapping system, and network entities (e.g. LISP tunnel routers) retrieve and cache only locally relevant state information on demand. Furthermore, the literature shows that the mapping system internals can be designed to be scalable [11].

Core-Edge Split: Casado *et al.* [9] analyze the main shortcomings of existing SDN architectures and point to the Fabric architecture as a solution. Fabric is based on an element called network fabric, a set of forwarding elements whose main function is packet forwarding. By taking base on this concept, they split the network into three components: hosts, edge switches, and core fabric. With this, rich network services such as isolation, mobility, or security are performed at the edge while fabric control is only responsible for packet forwarding. It is simple to establish a bijective relationship from Fabric components to LISP elements: tunnel routers perform edge switch functions, hosts are located on the EID space, and the core fabric corresponds to the RLOC space. From an abstract point of view, LISP offers an equivalent architecture to the one proposed by Fabric.

LISP SDN Building Blocks

In this section we analyze how specific LISP architectural elements can be used as SDN building blocks to understand the technical advantages and disadvantages of LISP as an SDN solution.

Flexible Namespace

The main LISP specification assumes IPv4 and IPv6 as address families, but it is flexible enough to allow using any other address families (e.g. MAC addresses). The LISP canonical address format (LCAF) allows defining ad-hoc address types that can be used for any purpose on a LISP system.

The template to define this type of address follows a simple TLV format (type-length-value). With this format, it is possible to define any

address type, including nested addresses of the same or different type. There are several address types defined at the time of this writing: AS number, geo-coordinates, application data, NAT-traversal data, multicast info, and so on. As an example, geo-coordinates addresses are used to carry geographical information along with any other address.

In general, such addresses allow LISP to map from any kind of identifier to any kind of locator, which means that, from an abstract point of view, LISP can map from any namespace to another. This address agnosticism enables rich network state programmability and can help to ease the interoperability challenges of heterogeneous SDN deployments

Distributed Mapping Database

Interface: The interface to exchange information with the mapping system is standard and open, and all the mapping system internal elements are hidden behind this interface (see Fig. 2). This allows the LISP data-plane devices to remain agnostic of the mapping system internal implementation. Such decoupling was put into test when the LISP beta-network deployed on the Internet (lisp4.net) replaced the existing mapping system, based on BGP, to a new one based on DNS without interfering with any of the LISP data-plane elements.

Arbitrary Information: Using the LISP flexible addresses (LCAF) described in the previous section, the mappings can contain any arbitrary information and be read/written from the mapping system using a standard interface. An SDN system can take advantage of this feature to store the network state. This is similar to what Onix [10] does with its own distributed databases.

Onix is a well known wide-area SDN deployment that addresses the lack of a general SDN control platform that can provide network-wide management abstractions. Onix provides an infrastructure to manage network state, on top of which different control-plane applications can be implemented. To offer this, Onix deploys its own database system to keep the network state and relies on the OpenFlow protocol to communicate with the network devices. Onix takes care of keeping consistent and distributed this network state over all network elements. With a LISP deployment, Onix could take advantage of LISP capabilities to provide similar functionality. First, it could use the mapping system with flexible addresses to keep network state and policies instead of deploying its own database system. Second, it could automatically reflect this state on the actual network if the network devices directly pull these policies from the mapping system using the LISP protocol.

Internal Scalability: The internal architecture of a specific mapping system varies depending on the type of information it is expected to store. Figure 2 also shows how the mapping system can use different internal implementations.

A mapping system indexing common IP addresses benefits from a hierarchical structure, such as DNS. This is the approach followed by the delegated database tree (DDT) based on [11],

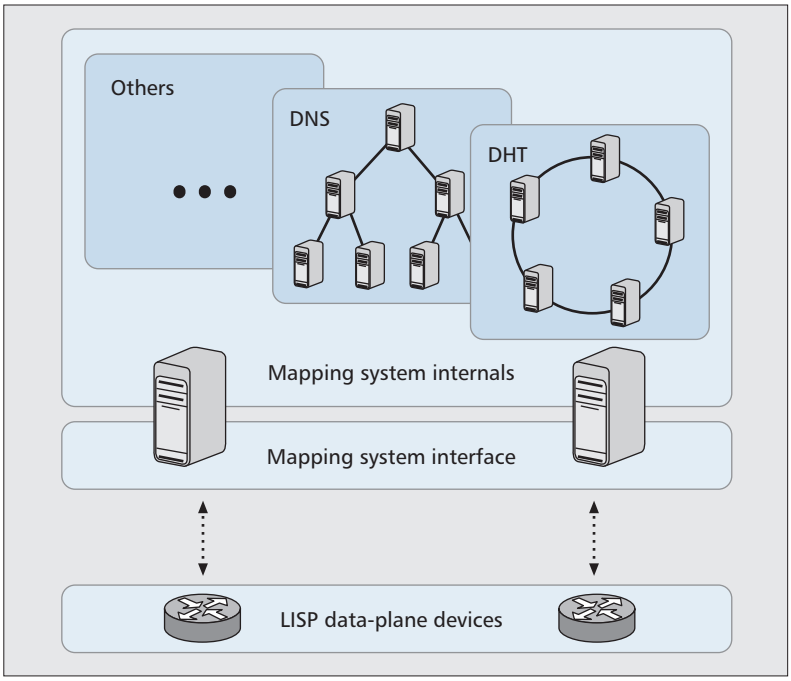


Figure 2. LISP mapping system.

the mapping system design used on the current LISP Internet deployment (lisp4.net). On the other hand, some deployments could require a flat name space; this is the case of non-aggregatable data such as character strings. For such requirements, a distributed hash table (DHT) design, rather than a DNS-like design, should be used. Although some initial efforts toward a DHT-like mapping system can be found in the literature [12], at the time of this writing only a hierarchical mapping system (DDT) has been successfully widely deployed.

Consistency: Levin *et al.* [7] expose the impact of a distributed SDN state on a logical centralized control application. While LISP still needs to deal with distributed trade-offs, its design allows mitigating them. The LISP mapping system is consistent and any snapshot of the distributed information reflects the desired control state. However, LISP network elements are eventually inconsistent, since an update on the mapping system is not instantaneously reflected on the data-plane. For instance, a LISP tunnel router can register new mapping information into the mapping system at any time, but an old version of the mapping can still be cached by remote tunnel routers. In order to minimize network inconsistency time, LISP defines two mechanisms to enforce up-to-date information at the data-plane. First, data-packets can carry an index of the current version of the mapping; second, a special control message can be used to explicitly notify remote parties of the mapping update.

Network Landmarks

Re-encapsulating tunnel routers (RTR) are special LISP tunnel routers that can be deployed on the RLOC space, rather than on the EID-RLOC edges. They receive LISP traffic, decapsulate it, look-up on the mapping system for the next hop,

re-encapsulate the traffic, and forward it. They give flexibility to the data path, offering network landmarks that data-packets can use.

These routers are a key element of a LISP SDN deployment. They can process the decapsulated traffic prior to re-encapsulating it again. This means, for instance, that traffic can be inspected, accounted, dropped, or modified at the re-encapsulating tunnel routers. An SDN approach can take advantage of these elements to set up network function devices. Devices such as firewalls, traffic analyzers, and accounting points, can be plugged, implemented, or virtualized on top of re-encapsulating devices. In that sense, Fig. 3 shows the abstract representation of a re-encapsulating tunnel router device with some network functions integrated that are used on demand.

Traffic Engineering

A mapping on the LISP mapping system can link an identifier to several locators. LISP allows defining a different priority and weight per locator. These values are used to specify the preference of the RLOCs to use to reach an EID as well as how to balance traffic among them. Besides that, LISP also introduces advanced traffic engineering capabilities by means of the explicit locator path (ELP). An explicit locator path is a list of hops through which packets have to be routed. The packets have to visit those locators in the same order as they are listed in the explicit path. These explicit paths serve as a mechanism to force traffic to follow a certain path on the locators space.

Priorities and weights also apply to locator paths, which means that an EID can map to several locator paths with different priority/weight attributes. Furthermore, such paths can be nested, creating sub-paths. This is done using EIDs instead of RLOCs as hops in the path. The final locator-only path will be obtained by a recursive look-up process. When a device finds that the next hop of the path is an EID, it will look-up on the mapping system to know the sub-path that this EID represents. Note that these sub-paths are subject to priority and weight values the same way as any other locator on the path. Using priority and weight, a LISP system can use different paths for the same destination where one path could be the most preferable while the others serve as backup, or many paths can be used at the same time to balance traffic.

Figure 4 shows an example. The traffic going to endpoint C should first go through M and N before being delivered at U. If that path is not available, then the traffic should be balanced in a 70/30 fashion over locators V and W. The traffic going to D should follow the path defined by α before reaching Z. As a backup, it can be also delivered directly on Z. In the example, α is used as a special identifier that represents a path instead of an endpoint.

Locator paths and re-encapsulating devices are tightly coupled, since in most of the cases the locator paths are used to force traffic to go through re-encapsulating devices. Explicit locator paths combined with re-encapsulating devices enable network programmability due to the ability to define custom programmable paths for packets in real time. Priority and weight parameters serve a fundamental role when deploying traffic rules. Traffic can be balanced among several paths and, thanks to recursion, to an arbitrary number of sub-paths. An SDN approach can deploy several re-encapsulating devices that also may implement (virtualized or not) network functions, and then program the mapping system to force the traffic to flow through these devices using explicit paths. Path nesting allows defining common sets of re-encapsulating devices that can be applied at once to specific traffic.

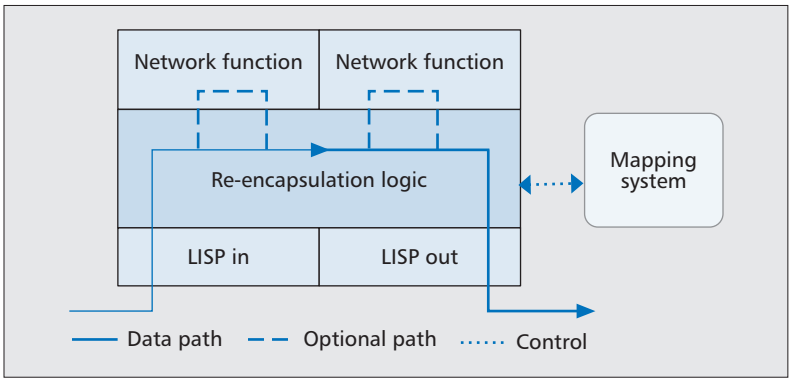


Figure 3. LISP re-encapsulating tunnel router.

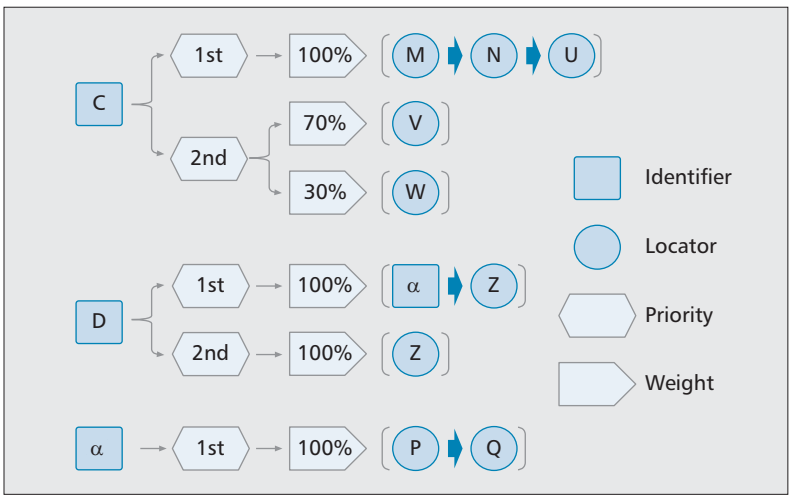


Figure 4. LISP traffic engineering.

Label System

Instance-ID is a 24 bit length identifier that can be associated with a certain EID. The identifier is included in the LISP header, and hence in all data-packets. Typically, this is used to carry VLAN tags or VPN identifiers. With this, network operators can split network policies and traffic, enabling multi-tenancy deployments.

However, instance-ID can be used beyond its original purpose. It is a 24 bit tag that can be appended to any data-packet to enable further features, not only multi-tenancy or address reusing. Specifically it can be used for routing scalability as well as management. In an SDN proposal such as Fabric [9], instance-ID can be used to tag flows that should be forwarded in the same way, simplifying forwarding on the core fabric and improving management and scalability.

LISP FOR SDN

Based on the previous analysis, this section discusses the advantages and drawbacks of applying LISP for SDN.

HIGHLIGHTS

Based on the analysis in the previous section, we highlight the most relevant features of LISP in SDN environments.

Scalability: As described previously, the mapping system provides scalability to the LISP system, an SDN solution can leverage this to provide a scalable network state database that can be directly queried by both data and control devices.

Interoperability: Given its flexible namespace and its label system, LISP is agnostic to the protocols it encapsulates and is well-suited to deploy overlays.

Inter-Domain: Network landmarks and LISP traffic engineering capabilities allow LISP to enforce policies on transit networks and make it suitable for inter-domain deployments.

BENEFITS

First, LISP has been designed to be incrementally deployable and to leverage current IP-based networks. Any existing IP-based network can incorporate common SDN features by simply upgrading some routers to LISP tunnel routers and connecting them to a mapping system.

Second, the shortcomings of traditional SDN protocols are motivating the emergence of hybrid SDN proposals that combine SDN with traditional network solutions [13]. Interestingly, due to its scalability and interoperability, LISP eases the deployment of the aforementioned hybrid SDN networks, specially since LISP can be incrementally deployed. Furthermore, thanks to its flexibility, LISP is well-suited to accommodate future protocols and new network approaches.

Finally, in contrast with common SDN protocols that are designed to operate mostly within a single domain, LISP allows SDN policies to be enforced across domains (e.g. DC-to-DC, DC-to-user's home). Well-placed LISP elements (e.g. re-encapsulating tunnel routers) make possible a programmable SDN deployment over a transit network (e.g. the Internet), something that is more complex to accomplish with traditional SDN protocols.

DRAWBACKS

Due to both how the protocol operates and its nature as a map-and-encap approach, LISP has some limitations that must be taken into account when considering LISP as a southbound SDN protocol.

Extra Headers: In order to encapsulate the traffic, LISP adds extra headers to the packets. This increments the packet size and reduces the available payload.

Mapping Resolution: LISP devices resolve and cache the mapping information on demand. The first packets of non-cached flows need to be either buffered or dropped until the mapping resolution process has been completed.

Mapping Updates: Any update on the mapping system is propagated over the network. However, this propagation involves some delay due to the signaling process, which can introduce latency in the system and/or produce packet losses.

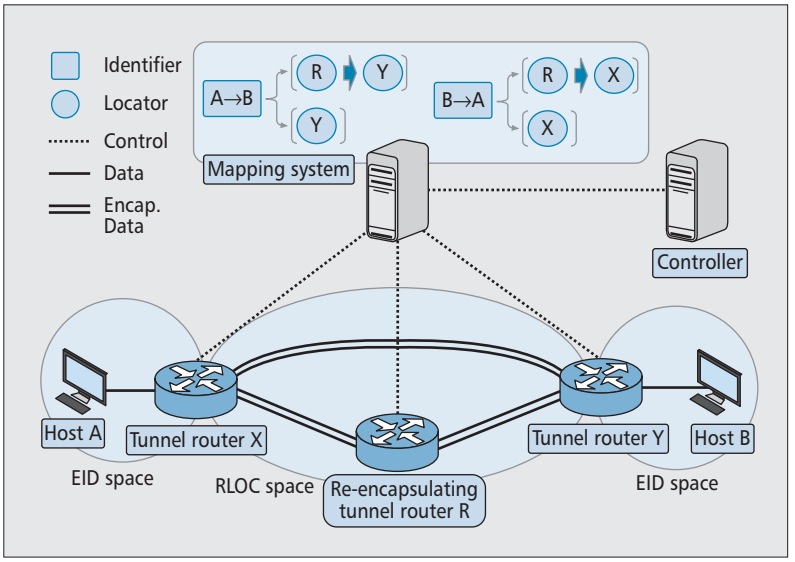


Figure 5. LISP SDN prototype.

Look-Up Support: While LISP defines how to convey different types of addresses in control messages, it does not define how to use all of those addresses to perform look-up operations.

Flat Data Support: Generally, mapping system implementations have been designed with hierarchical data in mind (e.g. IP addresses) and as such do not perform well when storing flat data (e.g. character strings).

Both *extra headers* and *mapping resolution* drawbacks are inherent in the LISP architecture. However, they do not have a strong impact on performance given that LISP encapsulation typically adds only 36 bytes (IPv4) or 56 bytes (IPv6) [1], and the LISP entities cache the mappings and because of the strong locality of traffic [14] achieve a hit-rate above 99 percent.

Regarding *flat data support*, the limitation can be solved with a DHT-based mapping system. Given that the interface to read/write mappings is open and standard, this limitation is not architectural and can be solved taking advantage of existing DHT databases.

To overcome the rest of the drawbacks we propose the following potential enhancements for the protocol above.

PROPOSED IMPROVEMENTS

The mapping updates limitation requires optimizing the mapping update signaling on SDN scenarios. In this context, we propose implementing a publish/subscribe mechanism for LISP mappings. The proposed mechanism is already being prototyped for the LISP project in OpenDaylight (opendaylight.org). The system operates as follows. Whenever a LISP data-plane device requests a mapping, the mapping system adds it to the list of subscribers for that mapping. Whenever the mapping data changes, the subscribers of that mapping are immediately notified, and thus they do not need to wait for the standard mapping update

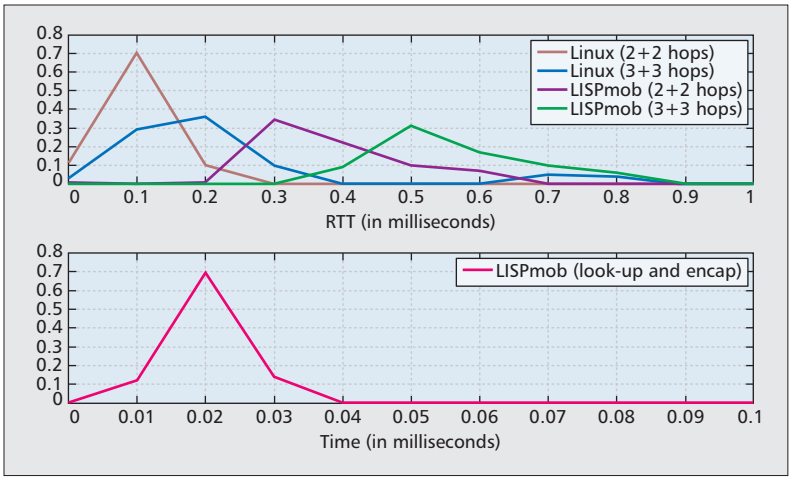


Figure 6. LISPmob induced delay.

propagation. The requester has to renew its subscription by explicitly requesting the mapping before a time-out. For scenarios where scalability and/or security is a concern, the subscription may be restricted to a set of predefined mappings or subscribers.

The *look-up support* needs to be extended beyond its current focus mostly in IP-prefixes. Most of the current SDN solutions operate the network in terms of flows. Traditionally, the minimal amount of information to identify a flow is its 5-tuple, even though normally in SDN more fields are used (e.g. OpenFlow). We advocate that LISP requires at least a look-up mechanism based on 5-tuples, despite the fact that in the future further look-up processes can be implemented, potentially leveraging on the OpenFlow tuple matching process.

Prototype

This section presents a prototype of a LISP-based SDN solution in order to validate its feasibility.

Setup

The prototype topology is depicted in Fig. 5. Two hosts (A and B) in different LISP sites are connected through the transit network via two tunnel routers (X and Y) and optionally via a re-encapsulating tunnel router. The mapping system stores mappings of source-destination EID tuples to RLOC space paths. These mappings are loaded by (borrowing OpenFlow terminology) a controller.

To implement the prototype, we instantiate a virtual machine running Linux for each of the elements on the topology. We connect the machines using virtual networks, emulating the topology depicted in the figure. On the machines that need LISP capabilities we run the open-source LISPmob implementation (lispmob.org) modified to support look-ups based on destination-source EID tuples.

On the described prototype we test two different scenarios, one where the traffic goes directly to its destination and another where the traffic goes through the detour introduced by the re-encapsulating tunnel router. We have a simple SDN application running on the controller that can dynamically set which path has higher priority.

Metrics

To extract relevant metrics we run 10 iterations injecting ping packet traffic during 10 secs per scenario (with and without detour) at a data-rate of 1000 pkts/s.

Packet Loss: The initial packet loss is due to the required mapping resolution signaling when we send a flow over a new path. We have measured an average initial packet loss of 3.0 packets dropped per iteration on the scenario without detour. This average packet loss goes up to 5.1 packets per iteration on the scenario with the extra LISP router due to the introduction of additional mapping resolution operations.

Delay: To measure how much delay is introduced by LISPmob we built an equivalent prototype without LISP capabilities, where traffic paths were configured modifying the routing tables on the Linux boxes. The top of Fig. 6 shows the PDF (probability distribution function) of the RTT (round trip time) for the scenarios considered. Note that without the detour round-trip traffic goes through four hops (i.e. two from A to B and two from B to A), while the detour introduces one extra hop in each direction (3+3) for a total of six hops. The bottom part of Fig. 6 shows how much time elapses from when LISPmob receives a new packet until it delivers the LISP encapsulated packet, that is, LISP look-up and encapsulation. The plots in Fig. 6 show that each LISP hop adds roughly 50 microseconds to the RTT, of which no more than 30 are due to LISP operations. The remaining latency is mostly due to the user ↔ kernel communication required by LISPmob. Nevertheless, the lookup and encapsulation operations may be optimized by router manufacturers to enable the performance of hardware implementations to be similar to that of traditional IP datagram forwarding.

Conclusions

In this article we have analyzed if LISP, in its current form, can be used for SDN. Our analysis concludes that the control-data decoupling, the network programability, and the centralized control enabled by traditional SDN solutions are already enabled by the LISP mapping system and supported by the rest of the LISP components. The major benefits of using LISP for SDN are that it keeps its incremental deployability and flexibility while providing scalability, interoperability, and inter-domain support, making LISP especially suitable for SDN deployments over legacy or transit networks, such as the Internet. However, despite its potential as an SDN enabler, there are some aspects of the protocol that should be extended to better fit the SDN use-case, mainly the signaling for the mapping updates and implementing support for an advanced look-up process. Finally, the presented prototype demonstrates that LISP is feasible for SDN scenarios.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments. 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] D. Farinacci *et al.*, "The Locator/ID Separation Protocol (LISP)," IETF RFC 6830, 2013.
[2] M. Jarsche *et al.*, "Interfaces, Attributes, and Use Cases: A Compass for SDN," *IEEE Commun. Mag.*, vol. 52, no. 6, June 2014, pp. 210–17.
[3] N. Foster *et al.*, "Languages for Software-Defined Networks," *IEEE Commun. Mag.*, vol. 51, no. 2, 2013, pp. 128-134.
[4] H. Kim and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Commun. Mag.*, vol. 51, no. 2, 2013, pp. 114–19.
[5] N. McKeown *et al.*, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, 2008, pp. 69–74.
[6] S. Sezer *et al.*, "Are We Ready for SDN? Implementation Challenges for Software-Defined Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, 2013, pp. 36–43.
[7] D. Levin *et al.*, "Logically Centralized?: State Distribution Trade-Offs in Software Defined Networks," *Proc. First Workshop on Hot Topics in Software Defined Networks*, ACM, 2012, pp. 1–6.
[8] S. H. Yeganeh *et al.*, "On Scalability of Software-Defined Networking," *IEEE Commun. Mag.*, vol. 51, no. 2, 2013, pp. 136–41.
[9] M. Casado *et al.*, "Fabric: A Retrospective on Evolving SDN," *Proc. First Workshop on Hot Topics in Software Defined Networks*, ACM, 2012, pp. 85–90.
[10] T. Koponen *et al.*, "Onix: A Distributed Control Platform for Large-Scale Production Networks," *OSDI*, vol. 10, 2010, pp. 1–6.
[11] L. Jakab *et al.*, "LISP-TREE: A DNS Hierarchy to Support the LISP Mapping System," *IEEE JSAC*, vol. 28, no. 8, 2010, pp. 1332–43.
[12] L. Mathy and L. Iannone, "LISP-DHT: Towards a DHT to Map Identifiers onto Locators," *Proc. 2008 ACM CoNEXT Conf.*, ACM, 2008.
[13] S. Vissicchio, L. Vanbever, and O. Bonaventure, "Opportunities and Research Challenges Of Hybrid Software Defined Networks," *SIGCOMM Comput. Commun. Rev.*, 44, 2, April 2014, pp. 70–75.
[14] F. Coras, A. Cabellos-Aparicio, and J. Domingo-Pascual, "An Analytical Model for the LISP Cache Size," *Proc. IFIP Networking*, 2012.

BIOGRAPHIES

ALBERTO RODRIGUEZ-NATAL received a BSc. (2010) and a MSc. (2012) in computer science from the University of Leon (Spain) and the Technical University of Catalonia (Spain), respectively. He is now a Ph.D. candidate at the Technical University of Catalonia and has been a visiting researcher at Cisco Systems (USA) and the National Institute of Informatics (Japan). His main research interests are future Internet architectures and Software-Defined Networking.

MARC PORTOLES-COMERAS received his degree in telecommunications engineering from the Technical University of Catalonia (UPC) and is currently working as a software engineer at Cisco Systems Inc., participating in the development of the LISP protocol architecture. Before joining Cisco he was a research engineer at the Centre Tecnológic de Telecomunicacions de Catalunya (CTTC) where he participated in multiple R&D projects. His current research interests are SDN and network virtualization solutions.

VINA ERMAGAN is a technical lead 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 SDN and network virtualization technologies ever since. She has initiated projects to implement LISP in Open vSwitch (OVS), OpenStack, and OpenDaylight. Vina received her MSc. in computer science from UC San Diego in 2008, and her BSc. in computer engineering from Sharif University of Technology.

DARREL LEWIS has more than 25 years of experience as an engineer for routing infrastructure vendors and network service providers. He has co-authored several LISP RFCs and he is currently a technical leader at Cisco Systems. Previously, he worked for Riverhead Networks as the lead consulting engineer. He is active in the North American Network Operators Group (NANOG), the Internet Engineering Task Force (IETF), and is a noted instructor in the fields of both Internet routing and security.

DINO FARINACCI is a technologist advancing the state of the art for the next-generation Internet. He was the original co-author for LISP dating back to 2007 and has the pleasure of writing two implementations of the protocol. He currently does consulting for large and startup networking vendors as well as users of such products. He is a software engineer by trade and a technology visionary by passion.

FABIO MAINO is a distinguished engineer at Cisco Systems, in the Chief of Technology and Architecture Office, where he leads the LISP research team. He has approximately 50 patents issued or filed with the US 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-APARICIO received a BSc. (2001), MSc. (2005), and Ph.D. (2008) degree in computer science from the Technical University of Catalonia (UPC), where he is now an assistant professor. In 2010 he joined the NaNoNetworking Center in Catalunya, where he is the Scientific Director. He has co-authored more than 15 journal and 40 conference papers. His main research interests are future architectures for the Internet and nano-scale communications.

Advertisers' Index

COMPANY	PAGE
Fraunhofer Heinrich Hertz Institute	5
ICC 2016 CFP	Cover 3
IEEE Member Digital Library	Cover 4
IEEE USA	8
Keysight	Cover 2, 1
Marconi Society	9
National Instruments	3

Currently Scheduled Topics

Topic	Issue Date	Manuscript Due Date
Ethics and Professionalism	November 2015	August 1, 2015
Wireless Communications, Networking, and Positioning with Unmanned Aerial Vehicles	May 2016	November 1, 2015
Bio-inspired Cyber Security for Communications and Networking	June 2016	November 1, 2015

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

IEEE International Conference on Communications
23-27 May 2016 // Kuala Lumpur // Malaysia
IEEE ICC'16: Communications for All Things

IEEE ICC 2016 CALL FOR PAPERS AND PROPOSALS

The 2016 IEEE International Conference on Communications (ICC) will be held from 23-27 May 2016 at Kuala Lumpur Convention Center, Malaysia, conveniently located in the middle of Southeast Asia, the region home to many of the world's largest ICT industries and research labs. Themed "Communications for All Things," this flagship conference of IEEE Communications Society will feature a comprehensive Technical Program including 13 Symposia and a number of Tutorials and Workshops. IEEE ICC 2016 will also include an attractive Industry Forum & Exhibition Program featuring keynote speakers, business and industry panels, and vendor exhibits.

TECHNICAL SYMPOSIA

We invite you to submit original technical papers in the following areas:

Symposium on Selected Areas in Communications

- Access Systems and Networks
Ahmed E. Kamal, Iowa State University, USA

- Cloud Communications and Networking
Dzmitry Kliazovich, University of Luxembourg, Luxembourg

- Communications for the Smart Grid
Lutz Lampe, University of British Columbia, Canada

- Data Storage
Edward Au, Huawei Technologies, Canada

- E-Health
Joel Rodrigues, University of Beira Interior, Portugal

- Internet of Things
Antonio Skarmeta, University of Murcia, Spain

- Satellite and Space Communications
Song Guo, University of Aizu, Japan

- Social Networking
Pan Hui, HKUST, Hong Kong

Ad-Hoc and Sensor Networks

Abdelhakim Hafid, University of Montreal, Canada
Cheng Li, Memorial University of Newfoundland, Canada
Pascal Lorenz, University of Haute-Alsace, France

Communication and Information System Security

Kejie Lu, University of Puerto Rico, Mayaguez, Puerto Rico
Yu Cheng, Illinois Institute of Technology, USA

Communications QoS, Reliability and Modelling

Kohei Shiomoto, NTT, Japan
Christos Verikoukis, CTTC, Spain
Charalabos Skianis, Aegean University, Greece

Cognitive Radio and Networks

Norman C. Beaulieu, BUPT, China
Lingyang Song, Peking University, China

Communications Software, Services and Multimedia Applications

Shingo Ata, Osaka City University, Japan
Fen Hou, University of Macau, China

Communication Theory

Marios Kountouris, Supelec, France
Marco Chiani, University of Bologna, Italy
Xu (Judy) Zhu, University of Liverpool, UK

Green Communications Systems and Networks

Sumei Sun, Institute for Infocomm Research, Singapore
Anura Jayasumana, Colorado State University, USA

Mobile and Wireless Networks

Adlen Ksentini, University of Rennes, France
Mohammed Atiquzzaman, University of Oklahoma, USA
Jalel Ben-Othman, University of Paris 13, France

Next Generation Networking and Internet

Rami Langar, University of Paris 6, France
Shiwen Mao, Auburn University, USA
Abdelhamid Mellouk, University of Paris-Est, France

Optical Networks and Systems

Walter Cerroni, University of Bologna, Italy
Krishna Sivalingam, IIT Madras, India

Signal Processing for Communications

Hsiao-Chun Wu, Louisiana State University, USA
Shaodan Ma, University of Macau, China
Tomohiko Taniguchi, Fujitsu Labs, Japan

Wireless Communications

Xiaohu Ge, Huazong University of Science and Technology, China
Dimitrie Popescu, Old Dominion University, USA
Hossam Hassanein, Queen's University, Canada
Rui Zhang, National University of Singapore

INDUSTRIAL FORUM AND EXHIBITION PROGRAM

IEEE ICC 2016 will feature several prominent keynote speakers, major business and technology forums, and a large number of vendor exhibits. Submit your proposals to the IF&E Chair.
Khaled B. Letaief (eekhaled@ee.ust.hk)

TUTORIALS

Proposals are invited for half- or full-day tutorials in all communication and networking topics. For enquiries, please contact Tutorial Program Co-Chairs.
Mike Devetsikiotis (mdevets@ncsu.edu)
Koichi Asatani (asatani@ieee.org)

WORKSHOPS

Proposals are invited for half- or full-day workshops in all communication and networking topics. For enquiries, please contact Workshop Program Co-Chairs.
Tarek El-Bawab (telbawab@ieee.org)
Fabrizio Granelli (granelli@disi.unitn.it)

ORGANIZING COMMITTEE

General Chair
Dato' Sri Jamaludin Ibrahim
CEO, Axiata Group, Malaysia

Executive Co-Chairs
Hikmet Sari
Supelec, France
Borhanuddin Mohd Ali
Universiti Putra, Malaysia

Technical Program Co-Chairs
Stefano Bregni
Politecnico di Milano, Italy
Nelson Fonseca
State University of Campinas, Brazil

Technical Program Vice-Chair
Jiang Linda Xie
University of North Carolina, Charlotte, USA

Industry Forums & Exhibition Chair
Khaled B. Letaief
Hong Kong University of Science and Technology, Hong Kong

Tutorial Program Co-Chairs
Mike Devetsikiotis
North Carolina State University, USA
Koichi Asatani
Kogakuin University, Japan

Workshop Program Co-Chairs
Tarek El-Bawab
Jackson State University, USA
Fabrizio Granelli
University of Trento, Italy

Conference Operations Chair
Hafizal Mohamad
MIMOS Berhad, Malaysia

Advisory Executive Vice-Chair
Datuk Hod Parman
Past Communication Commission General Director, Malaysia

Exhibition Chair
Nordin Ramli
MIMOS Berhad, Malaysia

IMPORTANT DATES

Paper Submissions:
16 October 2015

Tutorial Proposals:
13 November 2015

IF&E Proposals:
13 November 2015

Workshop Proposals:
17 July 2015

Paper Acceptance Notification:
29 January 2016

Camera-Ready Papers:
29 February 2016

Accepted and presented technical and workshop papers will be published in the IEEE ICC 2016 Conference Proceedings and submitted to IEEE Xplore®. See the website for author requirements of accepted authors. Full details of submission procedures are available at www.ieee-icc.org.