November 2014, Vol. 52, No. 11

IEEE Communications MAGAZINE

www.comsoc.org

The Future of Wi-Fi

- 5G Networks: End-to-End Architecture and Infrastructures
- Green Communications and Computing Networks

IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

IEEE Communications MAGAZINE

THANKS OUR CORPORATE SUPPORTERS

SAMSUNG

Anritsu
Test and Measurement Solutions

BEEcube

CISCO™

ROHDE&SCHWARZ

KEYSIGHT TECHNOLOGIES

November 2014, Vol. 52, No. 11

IEEE Communications MAGAZINE

www.comsoc.org

The Future of Wi-Fi

- 5G Networks: End-to-End Architecture and Infrastructures
- Green Communications and Computing Networks

IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

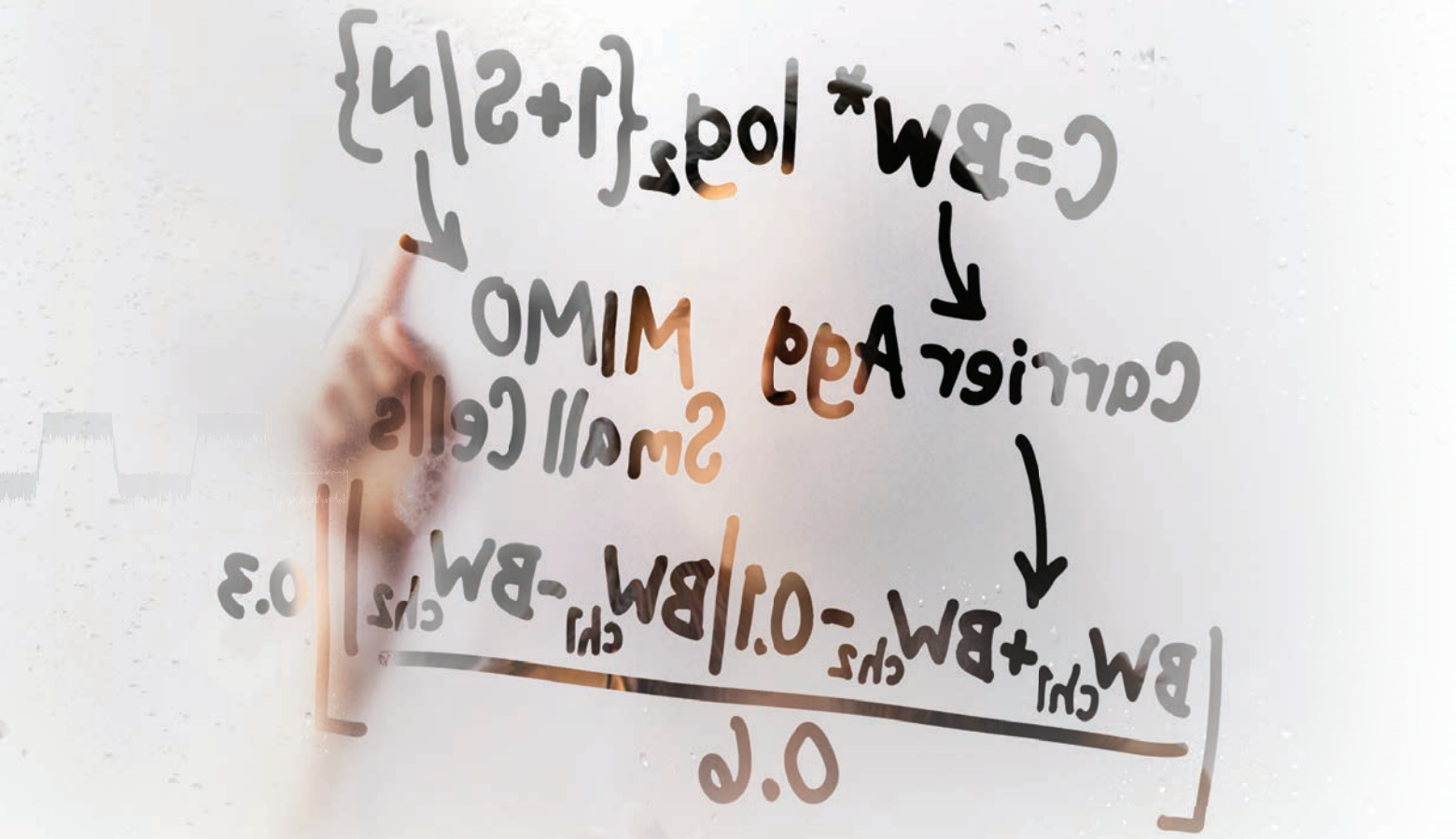
Eureka!

We'll help you get there.

Keysight 89600 VSA software

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

HARDWARE + SOFTWARE + PEOPLE = LTE-A INSIGHTS



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

USA: 800 829 4444 CAN: 877 894 4414

Keysight W1715EP SystemVue MIMO channel builder

Keysight N9040B UXA signal analyzer
with 89600 VSA software

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

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

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

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

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

HARDWARE + SOFTWARE

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

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

PEOPLE

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

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

KEYSIGHT TECHNOLOGIES

Unlocking Measurement Insights

IEEE Communications Magazine

NOVEMBER 2014, Vol. 52, No. 11

www.comsoc.org/commag

The Future of Wi-Fi

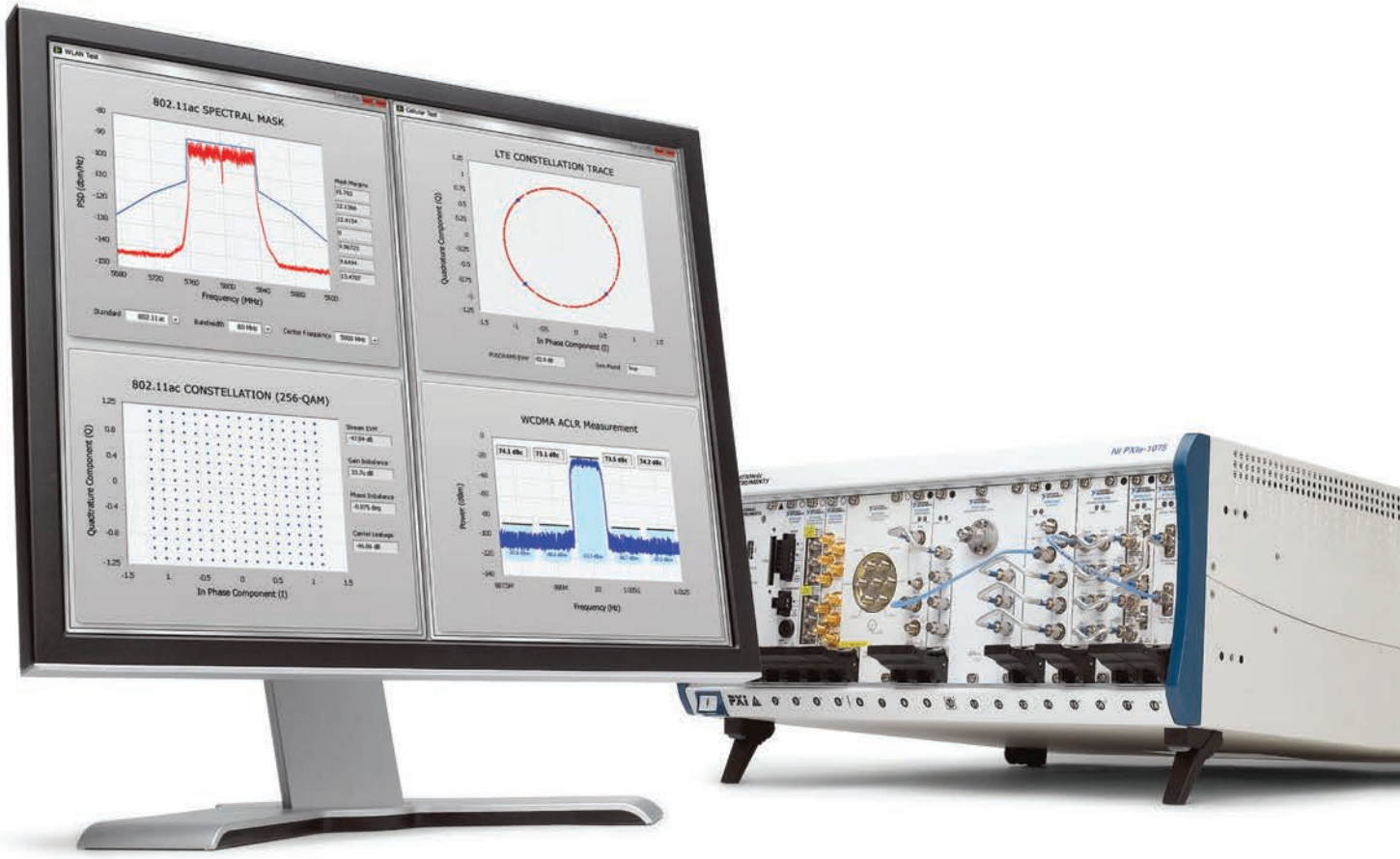
Guest Editors: Edward Au, Minho Cheong, Chiu Ngo, Carlos Cordeiro, and Weihua Zhuang

5G Networks: End-to-End Architecture and Infrastructure

Guest Editors: David Soldani, Kostas Pentikousis, Rahim Tafazolli, and Daniele Franceschini

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

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

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

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

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

GREEN COMMUNICATIONS AND COMPUTING NETWORKS

SERIES EDITORS: JINSONG WU, JOHN THOMPSON, HONGGANG ZHANG, AND DANIEL C. KILPER

ACCEPTED FROM OPEN CALL

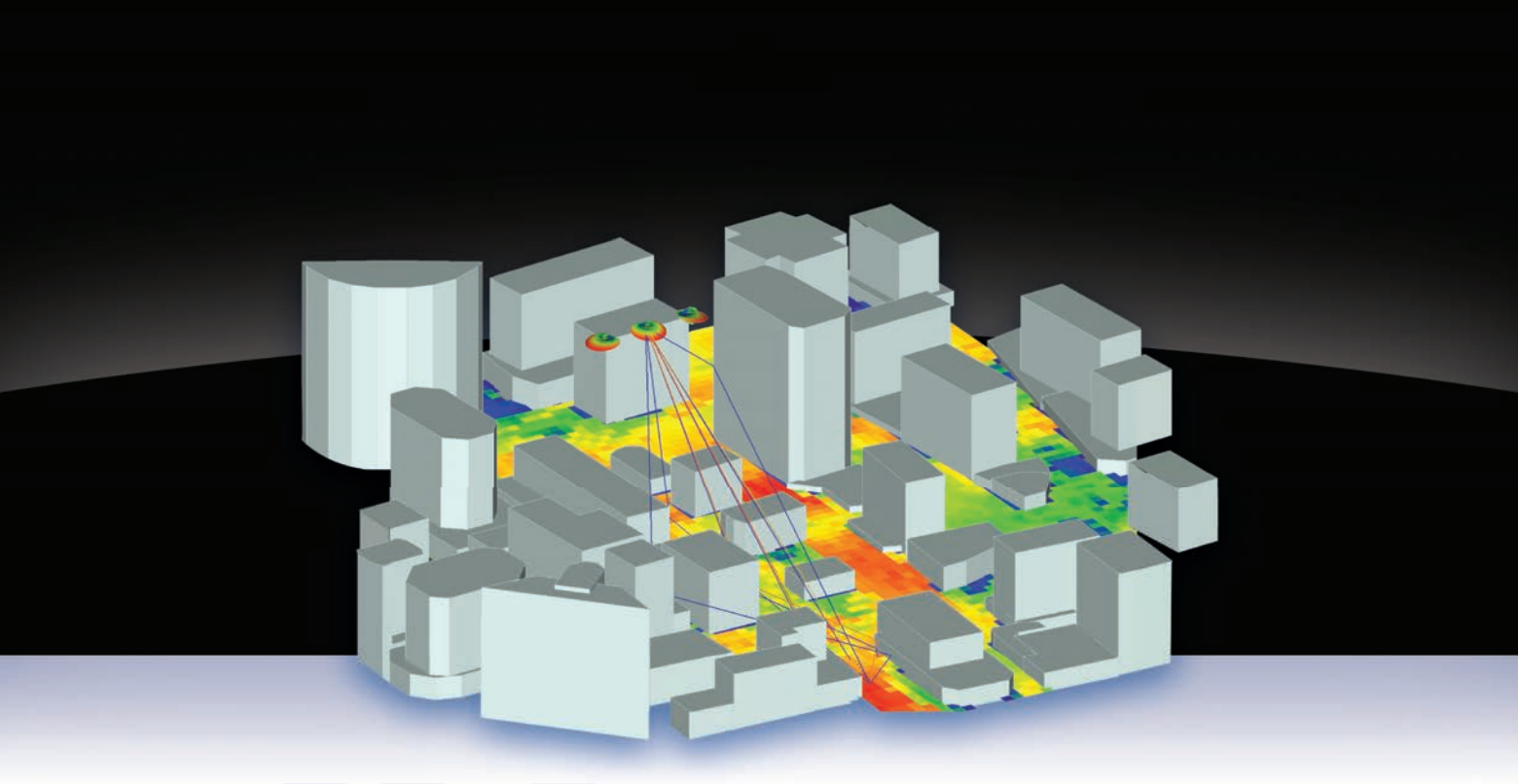
CURRENTLY SCHEDULED TOPICS

TOPIC	ISSUE DATE	MANUSCRIPT DUE
WIRELESS PHYSICAL LAYER SECURITY	JUNE 2015	NOVEMBER 15, 2014
5G SPECTRUM: ENABLING THE FUTURE MOBILE LANDSCAPE	JULY 2015	DECEMBER 1, 2014
UNDERWATER WIRELESS COMMUNICATIONS AND NETWORKS: THEORY AND APPLICATIONS	JULY 2015	DECEMBER 15, 2014
COMMUNICATIONS EDUCATION AND TRAINING: INDUSTRY CERTIFICATION AND UNIVERSITY ACCREDITATION PROGRAMS	MAY 2015	JANUARY 1, 2015
INTERNET OF THINGS/M2M FROM RESEARCH TO STANDARDS: THE NEXT STEPS	AUGUST 2015	JANUARY 15, 2015
SOFTWARE DEFINED 5G NETWORKS FOR ANYTHING AS A SERVICE	SEPTEMBER 2015	JANUARY 15, 2015
SOFTWARE DEFINED RADIO: 20 YEARS LATER	SEPTEMBER 2015	MARCH 1, 2015

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

TOPICS PLANNED FOR THE DECEMBER ISSUE

COMMUNICATIONS STANDARDS SUPPLEMENT
DISASTER RESILIENCE IN COMMUNICATIONS NETWORKS
USER-CENTRIC NETWORKING AND SERVICES
COMMUNICATIONS EDUCATION AND TRAINING
AUTOMOTIVE NETWORKING AND APPLICATIONS
CONSUMER COMMUNICATIONS AND NETWORKING
RADIO COMMUNICATIONS



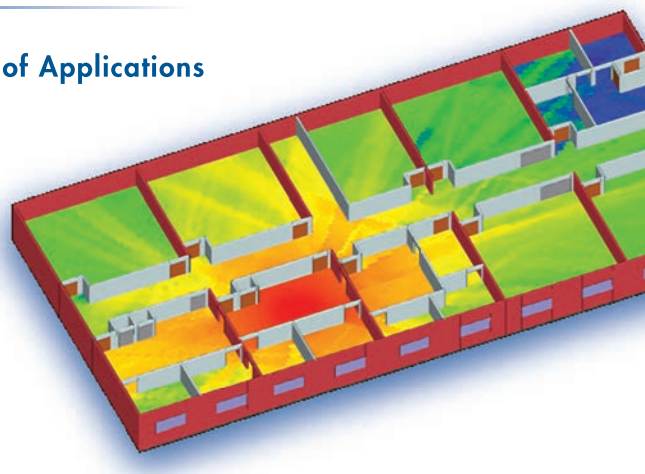
Remcom's Wireless InSite®

Radio Propagation Software for Wireless Communication Planning

Wireless InSite is a suite of ray-tracing models for analyzing EM propagation and communication channel characteristics in complex urban, indoor, rural and mixed path environments.

Wireless EM Propagation Capabilities for a Variety of Applications

- Indoor WiFi
- Moving vehicle or aircraft
- LTE and WiMax throughput analysis
- Tower placement for urban coverage
- Ad-hoc and temporary networks
- Base station coverage analysis
- Microcell coverage

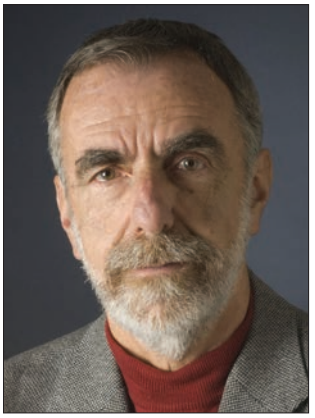


Now integrated with the Geospatial Data Abstraction Library (GDAL).
See all the latest enhancements at **www.remcom.com/wireless-insite-features**

REMCOM®

+1.888.7.REMCOM (US/CAN) | +1.814.861.1299 | **www.remcom.com**

IEEE AND COMSOC: ONE FOR ALL AND ALL FOR ONE!



SERGIO BENEDETTO

DOUG ZUCKERMAN

The IEEE Communications Society (ComSoc) is one of 45 technical societies and councils that provide the foundation for the IEEE's broad range of technical, professional, and humanitarian activities. These societies and councils are also part of a much broader, higher-level framework that includes Organizational Units (OUs) such as Member and Geographic Activities, Standards Activities, Educational Activities, etc. ComSoc and the extensive IEEE organization offer a wide range of valuable products and services to their members and the broader, global community. To succeed in fulfilling its mission, it is important that ComSoc recognize the wider IEEE landscape, and avoid "silo" thinking. This month's column, shared with Doug Zuckerman, provides perspectives on how the activities of ComSoc and its parent IEEE go hand in hand to enhance the value offered to IEEE and Society members by being, "One for All and All for One!"

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 Achieve- ment Award, and the Salah Aidarous Memorial Award.

GOING BEYOND THE SILO

Nowadays, the IEEE OUs holistically embrace going "beyond the silo" as a key paradigm for achieving their mission and thriving within the broader IEEE community. ComSoc formalized its approach to this through establishing an IEEE/ComSoc Coordination Committee to enhance the value of working with IEEE and its OUs. However, thinking beyond the silo was not always the case.

To understand the background, we refer back to the August 2010 *IEEE Communications Magazine* President's Page titled, "IEEE and ComSoc: Maximizing the Value":

"Many years ago, ComSoc enjoyed the strong support of the telecom companies that inhabited the pre-Divestiture world. It could count on scores of volunteers to organize and implement its key conference and publication activities. It was on a growth trajectory, with little need to look toward IEEE for 'help.' This set the stage for ComSoc to operate fairly independently, while still supporting its IEEE parent. Certainly, there were no proactive efforts aimed at strengthening the IEEE-ComSoc relationship beyond the minimal that was required.

"Time went by, the 1984 Divestiture resulted in fewer volunteers being supported by their companies to actively conduct ComSoc's activities, and both IEEE and the Society began relying more heavily on paid staff to assure continuation and growth in strong membership-oriented programs. ComSoc, as a very large IEEE society (second in size only to the Computer Society), began developing its own automated tools, including some that would aim at offering a 'portal' for all communications professionals' needs. At the same time, IEEE (ten times larger than ComSoc), was developing its own version of a 'portal.' The IEEE platform aimed then, as now, to meet the needs of not just one, but the full range of its operating units (OUs). Such OUs included all the technical societies falling under Technical Activities (TA), as well as (in today's terminology) the Member and Geographic Activities (MGA), Educational Activities (EA), Standards Activities (SA), IEEE-USA, and so on.

"The IEEE and ComSoc activities were no longer disjointed and had to be approached in concert. ComSoc's presidents and its division directors started actively participating on what had now become relevant IEEE boards and committees. Recognizing the benefits to both IEEE and ComSoc that could be attained by such engagement, the ComSoc Board of Governors established an IEEE/ComSoc Coordination Committee (ICCC).

"From the ComSoc Bylaws, 'This committee is responsible for the Society's internal coordination and cooperation with IEEE entities and for enhancing Society relations with IEEE governance and staff.' The committtee is chaired by a volunteer having experience at both the Society and IEEE level, and serves for a term concurrent with that of the Society president. Its membership includes the President (Byeong Gi Lee), Past President (Doug Zuckerman – also Chair) or President-Elect, VP Member Relations (Sergio Benedetto), IEEE Division III Director (Nim Cheung) and Past Director (Curtis Siller) or Director-Elect, and the Director of Sister & Related Societies (Roberto Saracco).

"It has currently been focusing on the following activities:

Information Sharing: IEEE/ComSoc Coordinating Committee members actively participate in various IEEE boards and committees, and share the results from and seek input to important initiatives, policy setting, strategic planning, and operational activities conducted by IEEE or its OUs. In addition to email, reports are provided at our key leadership meetings, including the Retreat, OpCom, and Board of Governors meetings.

Response to IEEE and OU Nomination Calls: IEEE and its OUs regularly seek potential candidates for a large number

of important leadership positions outside of ComSoc, ranging from committee membership (e.g. TAB Periodicals Committee) all the way up to IEEE President. Though ComSoc is not officially entitled to individual positions, its solid base of experienced and qualified volunteer leaders have much to offer (and bring back) when serving in the non-ComSoc positions at IEEE. Our coordination committee helps spread the word and encourages colleagues to be placed under such consideration by the IEEE, TAB, or other OU nomination committees.

Joint IEEE and ComSoc Staff Meetings: IEEE and ComSoc staff regularly meet from the Executive Director level on down to assure a common understanding of each organization's activities, to optimize their strategies and implementations. A goal is to provide the best possible value to the members while avoiding duplication of efforts within the corporate structure.

"These core committee activities have resulted in a significant – and growing – number of qualified ComSoc colleagues taking on increasingly important roles at the IEEE and other OU level, allowing ComSoc volunteer leadership to share their own experiences as well as learn from the experiences of others from outside other IEEE units."

COLLABORATION ON FUTURE TECHNOLOGIES AND APPLICATIONS

A fruitful area for "One for All and All for One!" collaboration is ComSoc's active engagement in programs and initiatives launched through the TAB Future Directions Committee (FDC) and the IEEE New Initiatives Committee (NIC), with support from ComSoc's Emerging Technologies Committee (ETC). These three committees are described below.

TAB Future Directions Committee (from the Technical Activities Board Operations Manual):

"The IEEE Future Directions Committee (FDC) reports to, and is a Standing Committee of, the TAB Strategic Planning Committee. The principal mission of the IEEE Future Directions Committee is to anticipate and determine the direction of existing, new and emerging technologies, and related issues, and to spearhead their investigation and development by IEEE in association with Societies/Councils (S/C's). The emphasis is on new, emerging technical areas either not adequately covered by existing S/C's, or which overlap the fields of interest of the existing S/C's, taking a collaborative view that would often involve other constituencies."

The Future Directions Committee works with the IEEE technical societies, such as ComSoc, to nimbly identify opportunities for IEEE to be a world leader in advancing new technologies for humanity. A streamlined process is in place to provide limited funding to get new initiatives off to a quick start.

An example is this year's launch of volunteer and staff activity on Software Defined Networks (SDN). ComSoc members active in the area worked with the FDC in identifying how IEEE could visibly contribute to development of this important technology. A special workshop was held by FDC to bring together experts with an interest in this field and who were enthusiastic about launching such a program through IEEE and its OUs, including ComSoc. To help enable IEEE future direction activities in SDN and other new topics, FDC has a small amount of seed funding available, currently limited at $US40K.

IEEE New Initiatives Committee (from the IEEE Bylaws):

"The IEEE New Initiatives Committee (NIC) shall be a committee of and report directly to the IEEE Board of Directors. The NIC shall identify, recommend, and monitor new initiative projects and programs consistent with IEEE's vision, mission, and Strategic Plan."

The New Initiatives Committee goes beyond just seeding new areas. It has substantial funding available for up to three years for a new technology initiative to be launched, operated, and transitioned into a sustainable program. It often will be a follow-on to initiatives seeded by the Future Directions Committee – FDC and NIC work very closely with each other.

An example of an IEEE New Initiative is the IEEE Cloud Computing Initiative (CCI; see cloudcomputing.ieee.org). NIC funded this initiative for three years starting in 2012. The main purpose of the initiative was to a) enable more efficient collaboration and achieve synergy across cloud computing activities that were scattered across IEEE's OUs, and b) show the world that IEEE is a key player in cloud computing (and networking). Its activities were initially structured into tracks covering Conferences, Publications, Education, Standards, and Testbed. A new track on Big Data was added around the third year, using the CCI infrastructure to nurture it toward becoming its own New Initiative with a three-year funding cycle starting in 2015. Multiple societies participated in CCI, with the Computer Society, Communications Society, Signal Processing Society, Vehicular Technology Society, and Consumer Electronics Society being the main contributors. Due to its size and complexity, IEEE assigned a program director to work closely with the volunteers and other staff to coordinate and manage the initiative's activities.

ComSoc Emerging Technologies Committee

To ease ComSoc's involvement with the TAB FDC and IEEE NIC initiatives and programs, ComSoc has a standing Emerging Technologies Committee. This committee nurtures potential new technical committees in hot new areas. For example, ETC created a sub-Technical Committee on Cloud Computing and Networking. This sub-TC has actively participated in the IEEE Cloud Computing Initiative by organizing panels and stimulating paper submissions to conferences and periodicals. ComSoc's Emerging Technologies Committee can also be a source of proposals for new initiatives in hot new technology areas.

BENEFITS OF COLLABORATION

The Cloud Computing Initiative is a good example of the benefits achievable through cross-OU collaboration. Each CCI track accrued benefits summarized as follows:

Conferences: Increased global visibility and higher attendance were achieved through CCI co-branding of existing or new IEEE (and some non-IEEE) conferences. Some examples were the new IEEE CloudNet, broadened IEEE COMPSAC, and the CCI-initiated and supported CCEM (Cloud Computing for Emerging Markets) conferences. Also, CCI organized regional cloud congresses in North America, Latin America, Europe, and Asia, built around existing or new conferences that covered cloud computing and networking. An example is the North American Regional Cloud Congress held at ComSoc's flagship IEEE GLOBECOM during the three-year Initiative funding period.

Periodicals: The Initiative brought together multiple Societies, including ComSoc, in launching new periodicals on Cloud Computing and Big Data. These included the *IEEE Transactions on Cloud Computing*, *IEEE Transactions on Big Data*, and *IEEE Cloud Computing Magazine*. Researchers and

others could now publish their work in high quality IEEE periodicals, and IEEE will receive additional revenue from Xplore and other access to these publications.

Education: In concert with IEEE Educational Activities, the Initiative developed extensive course material on cloud computing. The content is in modular form and will be offered in various combinations.

Standards: An important Initiative contribution of high value to industry was its development of two important standards on cloud interoperability: P2301 (cloud profiles) and P2301 (Intercloud). These still need to be proven in but show IEEE as a leader in addressing the growing challenge of many clouds having to interact with each other, i.e. "the Intercloud."

Testbed: The Initiative, through the IEEE Standards Association, engaged participation from across industry aimed at testing and refining the IEEE cloud interoperability standards for the Intercloud.

Big Data: The Cloud Computing Initiative provided a home for this increasingly important topic, which itself will become the IEEE Big Data Initiative starting in 2015. CCI facilitated launch of the new *IEEE Transactions on Big Data*, and provided opportunities for Big Data to be prominently included in conferences and IEEE periodicals. In addition, it worked with IEEE Corporate to establish an agreement with the New Jersey Big Data Alliance for future collaboration with IEEE.

The Cloud Computing Initiative is currently winding down as a NIC-subsidized initiative. However, plans are already in place for it to morph over to an IEEE Cloud Computing Program that will be centered in the IEEE Computer Society but which will nonetheless continue with active participation by at least a half dozen societies, councils, and other OUs (ComSoc plans to be a "partner" in this ongoing, sustainable program).

Examples of ComSoc engagement with other cross-OU initiatives or programs are those on Green ICT (which ComSoc leads, currently still subsidized by NIC) and Smart Grid (which the Power Engineering Society leads and which has already gone past its three-year NIC funding cycle).

THE CHALLENGE OF ONE FOR ALL AND ALL FOR ONE

"One for All and All for One!" is a worthy goal for one and for all. Achieving this goal is not automatic and faces challenges. Changes in industry structure and the world economy, together with "open access" trends, are presenting major challenges and opportunities to IEEE and its societies, including ComSoc. Collaboration across IEEE's technical societies, councils, and other OUs holds the key to IEEE's future in responding to these challenges. Areas impacting all of IEEE, along with some possible paths toward improvement, include:

- Budget Reform: Clarify and improve financial statement transparency to allow informed decisions.
- Simplify Rules: Simplify complex conference and publication rules to ease participation.
- Openness/Inclusiveness: Improve openness and inclusiveness in all our activities.
- Leadership Demographics: Engage more non-US volunteers in top leadership roles using nomination and balloting strategies.
- Industry Engagement: Establish "special interest groups" within TAB and the Societies/Councils to engage more industry.
- Standards Activities: Create a "standards committee" to focus and coordinate cross-Society/Council involvement with IEEE-Standards Association.
- Education: Leverage TAB's competences to grow Education into a major activity, comparable to Publications and Conferences.

ComSoc can and should play a key role in leading the way with its partner OUs in addressing these and other opportunities and challenges.

This month's column has highlighted the value provided by collaborating within the broader IEEE community. While it's likely easier for organizations to collectively go beyond silo thinking, it may be harder for individuals to do so. Each society, council, or other OU has its own unique culture and set of values. This needs to be taken into account as we leverage our unique strengths to achieve the vision of, "ComSoc for IEEE and IEEE for 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)

Dreamteam for success.

Signal generation and analysis for demanding requirements
When working at the cutting edge of technology, you shouldn't waste your time with inferior tools. Rely on measuring instruments evolved in the spirit of innovation and based on industry-leading expertise. Instruments like the R&S®SMW200A vector signal generator and the R&S®FSW signal and spectrum analyzer. Each is at the crest of today's possibilities. As a team, they open up new horizons.

See for yourself at www.rohde-schwarz.com/ad/highend

ROHDE&SCHWARZ

NURTURING EMERGING TECHNOLOGIES IN CLOUDS

ZHISHENG NIU, CHAIR, EMERGING TECHNOLOGIES COMMITTEE

Following the opening article by Prof. Andrea Goldsmith (former ETC chair), published in *IEEE Communication Magazine* in the December 2013 issue, I am very pleased to write the second edition of this column to further promote emerging technologies in the broad field of communications and related areas.

It is well known that communications are getting more ubiquitous and diverse and have become more interdisciplinary. Mobile computing, cloud computing, social networks, big data, and the Internet of Things have enabled us to access, use, and manage information almost anywhere and at any time. The communications market also began to evolve into a software and content driven market, where "communication" itself is getting more invisible or, in some sense, has been integrated into other fields like computing and social media. All these paradigm shifts have changed the rule of the game and, in fact, further expanded the scope of communications (and therefore the scope of emerging technologies). It is likely that we have been told to get our heads out of the clouds at one point in our lives. However, emerging technologies in the communications field are now forcing us to get our heads into the "cloud" and our hands on our mobile data. They have the potential to create new industries and transform existing ones at an ever-increasing rate.

While the emerging technologies are expected to drive the growth of the IT market, identifying them and determining how to best leverage them is not an easy job. This is mainly because emerging technologies are by their nature unknown, unproven, and risky, and are therefore difficult to manage. As a result, IT organizations (including ComSoc) are faced with the task of not only identifying relevant emerging technologies, but also developing their organizational awareness and motivation to nurture them. In this context, the emerging technologies committee (ETC) of ComSoc should play a key role in coordinating the wide range of activities in identifying and nurturing new technology directions within and also outside of ComSoc.

In this article I first outline the scope of two emerging technology subcommittees, which were not shown in the previous issue last year. Members with a common interest in these technology areas (including those in the subcommittees shown in the last issue) are strongly encouraged to join the subcommittees.

5G MOBILE WIRELESS INTERNET

The 5G "Mobile Wireless Internet" Technical Subcommittee will focus on exploring and elucidating all facets of the next generation of 5G Mobile Wireless Internet technologies, business and societal gaps, and challenges between the current 3G-4G-LTE access-only Internet models and the proper vision of 5G, evolutionary or revolutionary, to go beyond just access by embracing and facilitating the upfront integration of all new technologies (Internet of things, software-defined networks and network function virtualization, cloud computing, etc.) to be user-transparent, app-oriented, service-ready, ubiquitous, and lowest cost.

The objectives of this committee are to facilitate the worldwide harmonization of industry research and best practices for deployment user scenarios of the global 5G industry ecosystem, the built-in security and privacy by design in 5G, and explore the different ways to enable next generation Internet protocols over the next generation of empowered devices in order to reach convergence and end to end transparency.

This committee will also pursue a grander collaboration with IEEE TCs and non-IEEE industry standardization organizations as well as research enrichments from academia. For this purpose, it will invite at the next Globecom and ICC events 5G experts and IP designers from the IETF. This multi-discipline gathering of the members from this sub-TC will promote a common understanding to enable the convergence, governance, integration, and security of 5G.

SOFTWARE DEFINED NETWORKING AND NETWORK FUNCTION VIRTUALIZATION

The Software Defined Networks (SDN) and Network Function Virtualization (NFV) Technical Sub-committee will focus on exploring next generation networking technologies enabling software defined service delivery, network virtualization, network function virtualization, and the enablement of mobility. The subcommittee will analyze and drive integration around the touch points with all the other major IT inflexion points such as next generation IP, compute and storage virtualization, cloud, mobility, and the next generation applications. The key challenge to be addressed is to support multivendor networks in a software defined infrastructure that meets the demands of the next generation IT environments.

Topics addressed by the subcommittee will include network architecture, protocols and implementations that fully leverage the SDN/NFV concepts, strengths and weaknesses of current standards such as OpenFlow, alignment with cloud standards and IPv6 concepts, security considerations of SDN/NFV, innovative architectures, operations and service assurance in SDN/NFV-enabled environments; and education to develop the engineering talent needed to design, deploy, and operate SDN/NFV environments. This committee will harmonize its work with the Open Networking Foundation (ONF), IEEE, and non-IEEE organizations from academia and industry, including the academic research community, SDN/NFV and next-generation infrastructure projects, and standardization bodies.

Looking ahead, I would like to encourage all ComSoc members to broaden your vision across all the potential fields linked directly or indirectly to the discipline of communications and to propose new subcommittees accordingly. To do that, we can start from our ground breaking document "ComSoc Vision 2020", from where one can grasp the new trends in communications.

- ComSoc should better address multi-disciplinary areas in which communications has a role and include them in its technical scope. These could include borrowing concepts from biology, such as neuron networks, as a structural basis for coexistence of information processing and communications.
- ComSoc should address new areas, such as micro-communications.
- ComSoc should pay attention to content distribution networks, which have become very important for media and information distribution. The related areas of data mining, analysis and management, and information storage, accessibility, and security, should also be covered.
- ComSoc can be an important contributor to a 'smart city' vision and the

associated technologies, with recognition of social and economic considerations.

• ComSoc needs to become more heavily involved in sensor networks and clouds, which are going to be distributed all over the network, including edge networks, terminals, and communicating objects. Indeed, in many cases, they will be the network.

Members of the Society interested in organizing new technical groups should submit a written proposal to the VP of Technical Activities (VP-TA), copied to the ETC chair. The proposal should include the following sections:

• Name of proposed sub-committee.

• Charter of proposed subcommittee.

• Subcommittee website or plans to create a website (new subcommittees must create a website within one month of formation). The website must include the subcommittee name, charter, officers, and activities. The subcommittee will not be listed on the ETC website until it creates its own website.

• Rationale for organizing the subcommittee (i.e. differentiation of the proposed subcommittee relative to existing ETC subcommittees, ComSoc Technical Committees, and other professional groups).

• Proposed activities (e.g. conference and workshop organization, contributions to IEEE publications, etc.)

• Proposed officers (chair, vice-chair, secretary, and any other positions).

• Names and emails of at least 10 IEEE members that are members or interested in becoming members of the sub-committee.

The VP-TA will evaluate the proposal in consultation with the ETC chair and members of the Technical Activities Council and make a decision to approve or reject the formation of the sub-committee. The new sub-committee can then be organized with an approval and will report to the ETC chair. Each subcommittee will be required to submit an annual report to the ETC for review. Each subcommittee is also required to keep its website up to date. Subcommittees are also required to organize at least one meeting per year for all members. The meeting will ideally take place at an appropriate workshop or conference, but can be organized as a virtual meeting or conference call as well. A list of attendees to the meeting is required as part of the subcommittee annual report. The report, together with information on the subcommittee website, will be used to generate an annual evaluation by the ETC that will be sent to each subcommittee chair identifying strengths and making recommendations for improvement.

The VP-TA, together with the ETC chair and members of the Technical Affairs Council, will annually review the progress of each sub-committee and make one of the following decisions at the beginning of each calendar year.

• Continuation of the sub-committee (unconditional or under probation. Subcommittees under probation must address issues for which they are under probation within the following calendar year or they will be terminated.)

• Termination of the sub-committee. (Note: such decisions cannot not be made for sub-committees that have been formed for less than one year.)

• Recommend that the status of the sub-committee be changed to that of a full-fledged technical committee, in accordance with the Society's Constitution and Bylaws. A recommendation of this nature will be made to the VP-TA along with supporting documents, who will solicit approval of the new Technical Committee from the ComSoc BoG.

Last month I received a new proposal for establishing a new subcommittee on big data processing, analytics, and networking, which is under review. But of course, this is far less than enough. It is therefore highly encouraged to make more proposals in the coming years. Some potential areas include but not limited to:

1) Micro-communications (we are already looking at sensors.) In the coming years we will see a trend toward zero consumption and that will require new approaches to communications.

2) Future network (incorporate computation, storage, and big data into the network).

3) Data center networking or on-chip networking.

4) Wireless network convergence (e.g. outdoor and indoor).

5) Biomedical communications (brain-machine interface, in-body networks, compliments e-Health, …)

In summary, as my predecessor did, I would like to encourage all ComSoc members to participate in ETC subcommittees that intersect with their interests, and to propose new ones associated with emerging technologies in the field of communications and related disciplines. I also encourage all readers to contact me if you have other ideas about how ComSoc can promote and participate in emerging technologies, which will help maintain its leadership and vision in the field of communications.

CONFERENCE CALENDAR

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

2014

NOVEMBER

IEEE SmartGridComm 2014 — 5th IEEE Int'l. Conference on Smart Grid Communications, 3–6 Nov.
Venice, Italy.
http://sgc2014.ieee-smartgridcomm.org/

IEEE LATINCOM 2014 — IEEE Latin-American Conference on Communications, 5–7 Nov.
Cartagena de Indias, Colombia
http://www.ieee-latincom.org/

IEEE CCW 2014 — 28th IEEE Annual Computer Communications Workshop, 6–7 Nov.
Philadelphia, PA.
http://www.kkant.net/CCW2014/

IEEE OnlineGreenComm — 2014 IEEE Conference on Green Communications, 12–14 Nov.
Online
http://www.ieee-onlinegreencomm.org/2014/

WD 2014 — IFIP Wireless Days 2014 Conference, 12–14 Nov.
Rio de Janeiro, Brazil.
http://www.wireless-days.org/

RNDM 2014 — 6th Int'l. Workshop on Reliable Networks Design and Modeling, 17–19 Nov.
Barcelona, Spain.
http://www.rndm.pl/2014/

CNSM 2014 — 10th Int'l. Conference on Network and Service Management, 17–21 Nov.
Rio de Janeiro, Brazil.
http://www.cnsm-conf.org/2014/

IEEE LATINCLOUD 2014 — IEEE Latin American Conference on Cloud Computing and Communications, 24–25 Nov.
Rio de Janeiro, Brazil.
http://www.ieee-latincloud.org/2014/

ATNAC 2014 — 2014 Australasian Telecommunication Networks and Applications Conference, 26–28 Nov.
Melbourne, Australia.
http://www.atnac.org/

DECEMBER

IEEE CAMAD 2014 — IEEE Int'l. Workshop on Computer-Aided Modeling Analysis and Design of Communication Links and Networks, 1–3 Dec.
Athens, Greece.
http://www.ieee-camad.org/

Communications Society portfolio events appear in bold colored print.

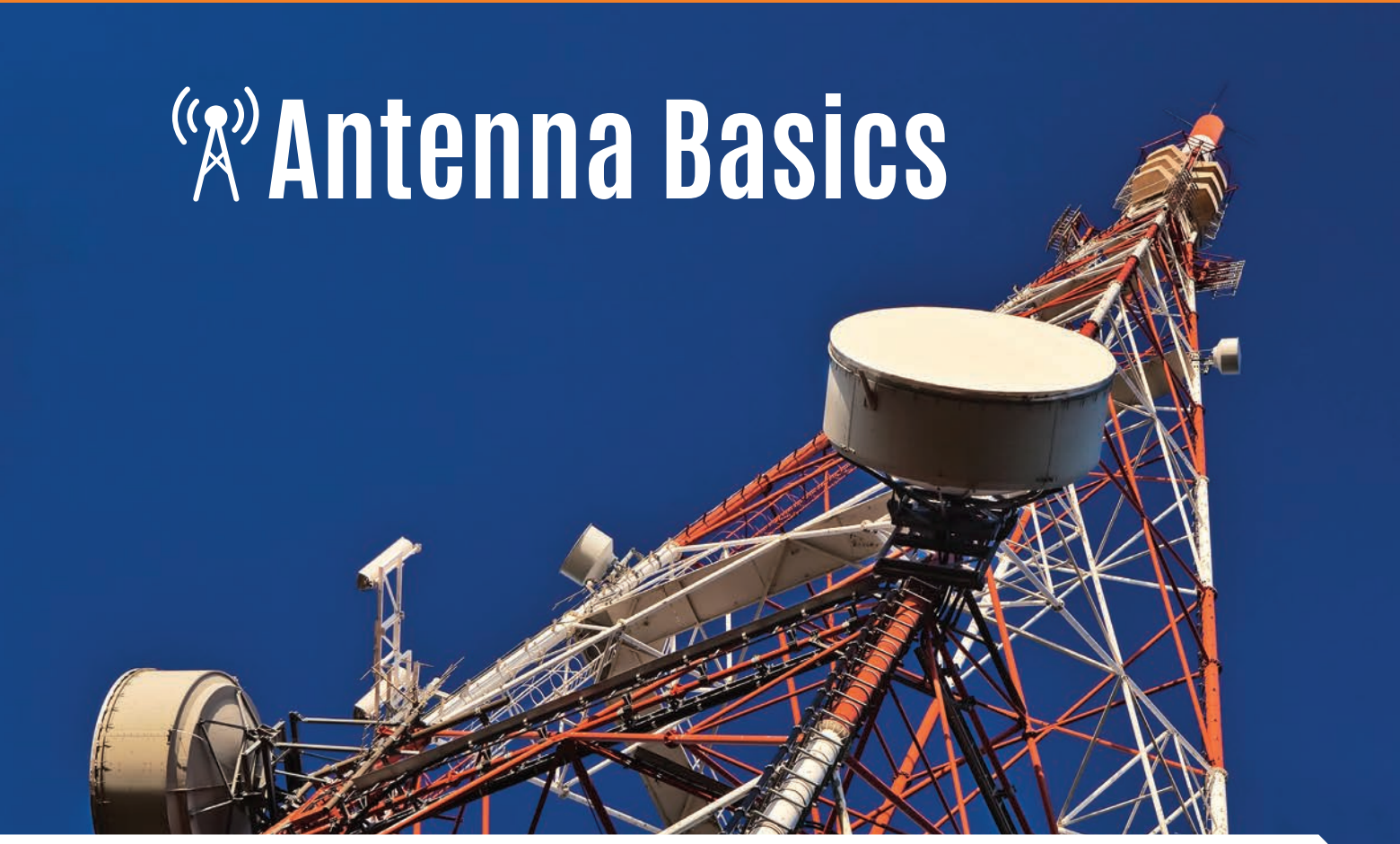
Communications Society technically co-sponsored conferences appear in black italic print.

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.

free Tutorials Now™

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

Antenna Basics



In this tutorial the basic functionality of an antenna is explained. Starting with Hertz's antenna model and a short introduction to the fundamentals of wave propagation, the important general characteristics of an antenna and its associated parameters (e.g. antenna gain, radiation pattern, bandwidth or VSWR) are defined. A more detailed review of the functionality of some selected antenna types (e.g. dipole or monopole) is also given.

Speaking will be Maik Reckeweg, Product Manager Antennas, Rohde & Schwarz GmbH, Munich, Germany, responsible for all the company's monitoring, measurement and communications antennas.

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

IEEE COMMUNICATIONS SOCIETY

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

FREE ACCESS SUPPORTED BY

ROHDE&SCHWARZ

NEW JERSEY INSTITUTE OF TECHNOLOGY PRESENTS EXCELLENCE IN RESEARCH LIFETIME ACHIEVEMENT AWARD TO YEHESKEL (ZEKE) BAR-NESS

BY RAYMOND L. PICKHOLTZ, COMSOC PRESIDENT 1990–1991,
AND CAROLE SWAIM, EXECUTIVE & VOLUNTEER SERVICES

Dr. Yeheskel Bar-Ness is a Distinguished Professor Emeritus of Electrical and Computer Engineering who has worked for four decades to advance the field of electrical and computer engineering.

At NJIT's seventh annual celebration of research excellence, October 2, 2014, the Board of Overseers honored Distinguished Professor Emeritus Yeheskel Bar-Ness for foundational contributions to the field of wireless communications. Bar-Ness received the 2014 Excellence in Research Lifetime Achievement Award for his groundbreaking work in electrical and computer engineering.

"Professor Bar-Ness is held in high esteem nationally and internationally by all of his peers. He has made very substantial contributions to basic knowledge in his area of expertise, and his work has influenced the quality of daily life in significant and positive ways," said NJIT President Joel Bloom.

Bar-Ness founded the Elisha Yegal Bar-Ness Center for Wireless Communications and Signal Processing Research at NJIT in 1985, which has long been at the forefront of wireless technology. The Center has contributed key technological advances in communications, including a set of algorithms that facilitate code-division multiple access, a digital cell phone technology that eliminates interference caused by high cell phone usage.

Bar-Ness and his colleagues have developed breakthrough technologies for industry, including a technology known as multiple input/multiple output, which uses antenna arrays to increase the bit rate of wireless communications. In other critical work, he led a collaborative project with Samsung to improve the capability of Worldwide Interoperability for Microwave Access (WiMAX), a certification mark for products that pass conformity tests established by IEEE, of which he is a Fellow and lifetime member.

"His central role in technological innovation has, without exaggeration, transformed the way we interact with each other every day. He has made major contributions to the evolution of wireless communications, technology basic to the practical development of systems that have made cell phones ubiquitous and which underlie the transport of data for myriad applications," said Don Sebastian, NJIT's senior vice president for technology and business development, and president and CEO of NJIT's New Jersey Innovation Institute. "His name appears on more than two dozen patents."

Professor Zeke Bar-Ness has been a tireless worker for the IEEE Communications Society over several decades. He has been active as a conference organizer, Editor, as well as in various leadership positions in Technical Committees.

However, Communications Society members particularly honor him for a singular accomplishment of lasting value to the Communications Society, namelythe establishment of *IEEE Communications Letters*, one of the archival journals published by the Society. In 1996 he conceived the idea for a journal that would allow researchers to publish brief papers in a short turnaround time. After a few meetings, he won the enthusiastic support of Steve Weinstein, who was then the Director of Publications. After approval by the Communications Society's Board of Governors, Zeke went to work immediately as the charter Editor of the new publication. The original plan was to publish the journal bimonthly in the first year; and, indeed, during that year six issues of *IEEE Communications Letters* were published. In those days, there were no automated web-based submissions and reviews, so everything was done by hand. Zeke personally read the submitted papers, found suitable editors, and at the end of each month decided on the organization of the articles for that particular issue. He served as its founding Editor for three years. *IEEE Communications Letters* is now a thriving and widely cited publication of the Communications Society.

Mark of Excellence: Presentation of NJIT's 2014 Excellence in Research Lifetime Achievement Award to Dr. Yeheskel Bar-Ness. Left to right: Philip L. Rinaldi, NJIT Overseer Chair and CEO of Philadelphia Energy Solutions, Zeke Bar-Ness, Joel S. Bloom, President of NJIT

In 2005 the Communications Society honored him with the IEEE Communications Society Publications Exemplary Service Award for his "outstanding, sustained, and visionary contributions to the Communications Society publication, *IEEE Transactions on Communications*, plus founder and first Editor-in-Chief of *IEEE Communications Letters*."

Zeke Bar-Ness has been a principal investigator or co-principal investigator on research grants or contracts supported by the National Science Foundation, the New Jersey Commission on Science and Technology, the U.S. Army, the U.S. Air Force, and the Naval Oceanic Center.

In 1973 he received the Kaplan Prize, awarded annually by the government of Israel to the 10 best technical contributors. Bar-Ness was named an Inventor of the Year in 2006 by the New Jersey Inventors Hall of Fame, sponsored by the Research & Development Council of New Jersey. He received his Ph.D. in applied mathematics from Brown University.

"Technology changes and I change with the technology," Bar-Ness said of his career in a recent biographical documentary.

His hard work, persistence, and can-do attitude are an inspiration for all of us.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

November 2014
ISSN 2374-1082

MEMBER RELATIONS

Asia/Pacific Region

Interview with Wanjiun Liao, Director of the Asia/Pacific Region

By Stefano Bregni, Vice-President for Member Relations, and Wanjiun Liao, Director of the Asia Pacific Region

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

In this issue, I interview Wanjiun Liao, Director of the Asia/Pacific Region. Wanjiun is the Y. Z. Hsu Scientific Chair Professor, a Distinguished Professor and the Department Chair of Electrical Engineering at the National Taiwan University (NTU), Taipei, Taiwan. She was an IEEE ComSoc Distinguished Lecturer (2011-2012), Associate Editor of *IEEE Transactions on Wireless Communications* (2003-2010), and of *IEEE Transactions on Multimedia* (2004-2007). She is a Fellow of IEEE. She has been a member of the IEEE Fellow Committee since 2013.

Stefano Bregni

Wanjiun Liao

It is my pleasure to interview Wanjiun and offer her this opportunity to present the activities of the AP Board.

Bregni: *Wanjiun, what outstanding characteristics of the Asia/Pacific Region would you highlight?*

Liao: The Asia/Pacific (AP) region, also known as Region 10, is the region with the largest number of members in IEEE ComSoc. The AP region is also one of the fastest growing economies with exciting opportunities. This further makes Region 10 the most popular region for ComSoc professionals to deliver Distinguished Lecturer Tours (DLT) and Distinguished Speaker Programs (DSP). In 2014 (as of Oct 1 2014) the total number of DLT/DSP in the AP region is 15, which is much higher than the total in the rest of the world!

Bregni: *What about the governance of the Asia/Pacific Region?*

Liao: The IEEE ComSoc Asia/Pacific Board (APB) is a well-organized and highly respected organization in ComSoc. The reputation results from the hard work of APB officers and volunteers over the years. The mission of APB is to address all ComSoc activities and programs related to AP members and chapters, including fostering provision of technical activities and information services to our members, expanding membership in AP, and reflecting the interests of AP members in ComSoc policies and procedures.

Bregni: *How is the Asia/Pacific Board organized?*

Liao: In the APB, there is one director, three vice directors, one treasurer, and one secretary, plus five operation committees, supported by the AP office in Singapore. We organize APB meetings, including the steering meeting and the general meeting, twice a year, at ICC and GLOBECOM. The steering meeting is to handle challenges in promoting APB activities and to explore new services to our members. The general meeting is to provide a good platform for APB members to make friends, share information, discover new research directions, and facilitate academic-industry collaboration.

Bregni: *What are the roles and responsibilities of the various APB Officers?*

Liao: In APB, the Director and the three Vice Directors oversee the provisioning of the APB activities to all ComSoc members in the AP region. The technical activities and service provisioning are organized into five committees, each with one chair plus several vice chairs: Technical Affairs Committee (TAC), Membership Development Committee (MDC), Information Service Committee (ISC), Meeting and Conference Committee (MCC), and Chapter Coordination Committee (CCC). The volunteers come from academia and industry, with a good mix of geographical locations, gender, and seniority.

Bregni: *It looks like a complex structure! Would you tell us more about the scope of those five Committees?*

Liao: The mission and plan of each of the five committees are summarized as follows:

1. TAC: to promote technical activities and to foster award activities, including AP young researcher, outstanding paper awards, and IEEE GOLD awards, for ComSoc members in the AP region.

2. MDC: to collaborate with ComSoc chapters in the AP region to promote academic and industry membership, and to liaise with sister and related societies for professional activities.

3. ISC: to create and distribute APB newsletters, to maintain the APB webpage, facebook, and to manage on-line DLT/DSP programs to AP region members.

4. MCC: to encourage AP region members to organize, host, and participate in AP regional and ComSoc flagship conferences.

5. CCC: to coordinate with IEEE ComSoc Membership Development Program (MDP) to run DLT/DSP programs and to manage chapter activities in the AP region.

Bregni: *What are your plans in the short term?*

Liao: In the future we will continue our good tradition, address the new needs of our members, and tackle new challenges. I believe we have formed an excellent team of APB officers who have already done a very good job. We do need more involvement and strong support from our members. There are many opportunities for ComSoc members in the AP region in the near future, but to best exploit these opportunities we AP region members must work together!

Distinguished Lecturer Tour of Xiaoming Fu in Shanghai and Beijing, China, May 2014

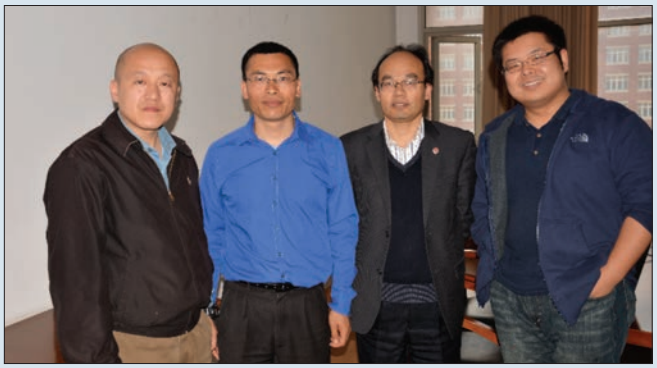
By Xiaoming Fu, Univ. of Goettingen, Germany

Having the opportunities to deliver several lectures on data networking and Internet computing in Asia and Pacific, Europe, North America and South America, I have been honored to be appointed as an IEEE Communications Society Distinguished Lecturer in January 2014.

China is my home country where I grew up and was educated. In recent years I have participated in several EU-sponsored or bilateral projects involving some Chinese partners, which provided much enjoyment and fresh experiences with different Chinese colleagues. Nevertheless, my personal feeling is that they are more project-driven, unlike a global organization such as IEEE ComSoc, which could bring a different value for the whole research community.

Upon the invitations of the IEEE ComSoc Shanghai Chapter and the Beijing Chapter, I had the opportunity to make my first DLT to China. IEEE ComSoc Asia-Pacific Project/Admin Executive Ewell Tan was extremely efficient and professional in communicating with local chapters and sectors on my DLT schedule. The initial thought was to deliver lectures in Nanjing, Shanghai, Wuhan, and Beijing, but due to time conflicts I could not visit Wuhan, and some of my lectures ("Fine-Grained Multi-Resource Scheduling in Cloud Datacenters" at Nanjing University in Nanjing and Fudan University in Shanghai, as well as "Content Distribution: from Client/Server to Content-Oriented Publish/Subscribe System" at Tsinghua University in Beijing) were delivered outside the DLT program due to synchronization issues with the local chapters. Although these lectures were not directly under the IEEE DLT flag, Ewell encouraged me to deliver them as planned, which actually turned out to be well perceived by colleagues and students in these institutions.

After my Nanjing trip, I arrived in Shanghai on 4 May and gave a lecture at Fudan University on the morning of 5 May. After lunch I visited Shanghai Jiao Tong University (SJTU), one of the top engineering universities in China, for the first time, and gave my first DLT lecture on the design, implementation, and evaluation of scalable microblogging systems. There were over 30 attendees, some faculty members, and many graduate and undergraduate students from SEIEE, including all students from a seminar course usually planed for that slot. The audience showed a great interest in the topic, and I enjoyed the interactions during the talk and in the Q&A period after the presentation. In particular, I received emails from some students after the seminar. Some of them wanted to study in Göttingen and work on related topics. My hosts Prof. Xinwan Li (IEEE ComSoc Shanghai Chapter Chair, and Vice Dean of SJTU-Michigan University Joint Institute), Prof. Xinbing Wang (IEEE ComSoc Shanghai Chapter Chair), and Dr. Xiaohua Tian were very thoughtful, and warmly invited to visit their labs in SEIEE as well as the Michigan University-SJTU Joint Institute. From my personal feeling, SJTU is certainly China's most modern university, owing to their broad visions, maximal elimination of bureaucracy, and efficient adoption of a western educational culture. SJTU also has an open mind to attract world-class experts and scientists to work there or collaborate with them. It's amazing that many faculty members hold U.S., Canadian, Japanese, or European Ph.D. degrees, and they are implementing the tenure-track system with competitive salary and expectations. SJTU's Minhang campus is huge, beautiful, and elegant. I did not have the time to visit it except for the office buildings of Prof. Li, Prof. Wang, and an in-campus coffee bar, but I definitely want to see the entire campus on a future trip.



Xinbing Wang, Xiaoming Fu, Xinwan Li, and Xiaohua Tian at SJTU (from left to right), right after the Q&A.

Lin Zhang, Tarik Taleb, Dieter Hogrefe, Xiaoming Fu, and Qimei Cui at BUPT campus.

DLT at ICT-CAS.

My next DLT lecture took place at Beijing University of Posts and Telecommunications (BUPT) on 7 May 2014. In the classroom I was delighted to meet another DLT lecturer, Dr. Tarik Taleb, a colleague from NEC Europe Networking Lab, Heidelberg, Germany, who gave a lecture just before mine. We have known each other for quite some time and we are both located in Germany. His lecture focused on cloud computing architecture perspectives, while mine was more focused on the Internet service and systems point of view. There were roughly 40 attendees, mostly faculty members and graduate students from the Communications Engineering department, from which I saw there was a keen interest in learning about the social networks domain and other emerging Internet services. After the lecture I also had the pleasure of meeting our current EU project partners (Prof. Lin

(Continued on Newsletter page 4)

CONFERENCE REPORT

Fifth International Conference on Wireless Communications and Signal Processing (WCSP, 2013), Hangzhou, China

By Zhaoyang Zhang, TPC Co-Chair, Zhejiang Univ.; Caijun Zhong, Zhejiang Univ.; Guangguo Bi, Chair of Nanjing Chapter

The 5th event in the successful series of WCSP, 2013 International Conference on Wireless Communications and Signal Processing (WCSP 2013) was held 24-26 October at the Dragon Hotel, around the picturesque West Lake in the beautiful city Hangzhou, China. The conference attracted more than 250 academics, engineers, and students from 19 countries and regions around the world.

WCSP 2013 was co-sponsored by the IEEE ComSoc Nanjing Chapter and the IEEE SP Society Nanjing Chapter. It was also technically co-sponsored by the IEEE Communications Society and the China Institute of Communications. WCSP 2013 was hosted by Zhejiang University, China, and co-organized by Southeast University, China, Nanjing University of Posts and Telecom., China, PLA University of Science and Tech., China, and the University of Science and Tech. of China.

Following the great success of WCSP 2009, WCSP 2010, WCSP 2011, and WCSP 2012, WCSP 2013 aims to bring together international researchers from academia and practitioners from industry to meet and exchange ideas and recent research advances on all aspects of wireless communications and signal processing. This year WCSP 2013 received a total of 601 submissions from 27 countries and regions around the world. All papers were rigorously and independently peer-reviewed by more than 350 specialists from universities, institutes, and companies around the world. Based on the level of relevance, originality, technical contributions, and presentation quality, 262 high quality papers were selected for publication in the final conference proceedings, yielding an acceptance ratio of 43.59%. Accepted papers were organized into 46 sessions with four technical symposia: Communication Theory Symposium, Wireless Communications Symposium, Signal Processing for Communications Symposium, and Wireless Networking Symposium.

In addition to the exciting technical sessions, the technical program of WCSP 2013 also featured six exceptional and splendid keynote speeches and an invited special talk delivered by distinguished experts from the IEEE Communication Society. The keynote speeches addressed the following topics: "Towards a theory of security for wireless networking" by Prof. P. R. Kumar, IEEE Fellow, Texas A&M University, USA; "Network localization and navigation" by Prof. Moe Win, IEEE Fellow, Massachusetts Institute of Technology, USA; "Relay-by-smartphone: a dual mode ad hoc network system for disaster-affected areas" by Prof. Nei Kato, IEEE Fellow, Tohoku University, Japan; "A new paradigm for mobile social networking: social tie, group utility maximization and privacy" by Prof. Junshan Zhang, IEEE Fellow, Arizona State University, USA; "Cognitive wireless networks: enabling 5G mobile communications" by Prof. Ying Chang Liang, IEEE Fellow, Institute for Infocomm Research, Singapore; and "Challenge of signal processing in 5G wireless" by Mr. Ganghua Yang, Huawei Technologies Co., Ltd., China. The invited special talk entitled "What's next after OFDM?" was given by Prof. Xianggen Xia, IEEE Fellow, University of Delaware, USA. All talks aroused much interest and received high praise from the attendees.

Best Paper Awards were given to 10 researchers' work covering various aspects of wireless communications and signal processing, namely: "Community detection based reference points clustering for indoor localization in WLAN"; "Secrecy-based channel assignment for device-to-device communication: an auction approach"; "Precoder design for dual-stream MIMO multicasting"; "DCT-based channel estimator for OFDM systems: threshold setting and leakage estimation"; "On the optimum energy efficiency for flat-fading channels with rate-dependent circuit power: time-invariant case"; "Recursive geometric water-filling for wireless links with hybrid energy systems"; "Ergodic capacity analysis of dual-hop ZF/MRT relaying systems with co-channel interference"; "Power allocation strategy for MIMO broadcast channels with receiver cooperation"; "Relay selection scheme in the presence of co-channel interference"; and "Optimal cooperative sensing and resource allocation in cognitive radio networks".

The General Co-Chair, vice president of Zhejiang Univ., Prof. Zhaohui Wu gave the welcome speech at the opening ceremony.

Prof. P. R. Kumar delivered the keynote talk .

The winners of the Best Paper Award at the Banquet.

WCSP 2013 also offered participants the opportunity to visit the beautiful city of Hangzhou and its surroundings, to enjoy the wonderful , romantic, big live-action performance 'Impression West Lake', and to taste the special Hangzhou dishes as well.

Acclaimed by the participants for its high quality technical program and excellent organization and local arrangements, WCSP 2013 achieved a great success. All 262 papers included in the conference proceedings are included in IEEE Xplore and indexed by EI Compendex. We believe that, under the guidance of the steering committee and with the continuing efforts from all the

(Continued on Newsletter page 4)

CONFERENCE REPORT

Looking Back at the Development of our Technologies

Dr. Jacob Baal-Schem – SLM, Tel-Aviv University, Israel

Due to the pace of development of communications technologies, we hardly have time and interest to turn back and look at our achievements, and to guide our students accordingly.

Just think that:

- 100 years ago there were no official broadcasting stations in the world.
- 50 years ago you could hardly find a portable radio set and all receivers worked on vacuum tubes.
- 40 years ago a "walkie-talkie" was a huge and heavy piece of equipment.
- 20 years ago the fastest way of written communication was by facsimile.

Look around and you can still meet the people who have developed the technologies that are so common to all of us today. Look around and you can sense the social impacts of these technologies on our daily life. This gives us a special opportunity to learn and discuss with those "elders" how these technologies were developed and receive lessons "from the mouth of the horse."

Actually, we seldom compare the long evolution of music recordings — 78rpm vinyl records to the MP3 and MP4 files of our children — and think how much effort was involved and how many scientists and technicians have spent their lives to achieve these changes. This is especially true for young students and even for young engineers. The past seems unimportant to many of them, while actually, "Who controls the past controls the future. Who controls the present controls the past" (George Orwell in "1984").

The IEEE Israel Section (actually, members of the ComSoc Chapter) has initiated, and IEEE Region 8 has approved, organizing a series of HISTory of ELectrotechnology CONferences – HISTELCON. The first HISTELCON was held in Paris, France, in 2008. The second was hosted by the Spain Section in Madrid, Spain, in 2010, and the third was hosted by the Italy Section and the local University in Pavia, Italy, in 2012. Approaching the 2012 event, the IEEE History Committee decided to cancel its bi-annual History Conference and all Committee Members participated in HISTELCON. The next HISTELCON is planned to be held at Tel-Aviv University, Israel in late August 2015, jointly with IEEE History Committee and Center and with a new partner: the International Committee on History of Technology (ICOHTEC). There is already interest to hold a future HISTELCON in Tokyo, Japan, and there is hope that the HISTELCON series will become the flagship IEEE Conference on the History of Technology.

Each of these Conferences brings together scientists, technologists, and historians to consider the ways by which our technologies developed, mainly during the 20th century. They consist of keynote lectures by eminent scientists, frontal presentations by the participants (based on reviewed abstracts), panel discussions, as well as visits to technical sites and the participation in social events, during a full week of activities.

The technology who's history is discussed incorporates state-of-the-art knowledge and is valued for its inventiveness and wide socio-cultural implications. As such, these are the focal points for research by historians of technology, scientists, and engineers exploring the emergence of their own field of expertise, as well as for economists, sociologists, and others.

The organizers aspire to a multifaceted picture of the developments of such technologies from various approaches, with talks discussing subjects that include (but are not restricted to) the origins, evolution, and demise of various techniques and methods, their employment, spread, and appropriation, the cultural, social, military, economic, scientific, natural, and technical factors that shaped these events, and the ways by which technologies influenced societies that adopted them.

The IEEE Israel Communications Chapter, the first Chapter of the first Section in Region 8, has enthusiastically participated in the organization and in the program of these Conferences, and all ComSoc members are heartily welcomed to participate in HISTELCON. For any questions, please contact: j.baal.schem@ieee.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

JACOB BAAL-SCHEM, ISRAEL (J.BAAL.SCHEM@IEEE.ORG)
EWELL TAN, SINGAPORE (EWELL.TAN@IEEE.ORG)

A publication of the IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

DISTINGUISHED LECTURER TOUR/*Continued from page 2*

Zhang and Prof. Jun Guo) and IEEE ComSoc Beijing Chapter Chair Prof. Xiaofeng Tao, and discussed some issues related to crowdsourcing and mobile cloud computing.

I delivered my third DLT lecture at the Institute of Computing Technology, Chinese Academy of Science (ICT-CAS), on 8 May 2014, with some 15 attendees, primarily the members and graduate students from Prof. Yiqing Zhou's lab. In the beginning Dr. Xue Han presented the structure of ICT-CAS and research activities in their lab on behalf of Prof. Yiqing Zhou, who could not attend the lecture due to sickness. My understanding is that they have primarily focused physical layer, link layer, and network layer functions in the 4G/5G direction, complementing what we have been pursuing in the network layer and above.

In summary, I had a pleasant and fruitful first DLT in China. I enjoyed it very much, thanks the generous support of IEEE ComSoc and also the hospitality of local chapters.

WCSP 2013/*Continued from page 3*

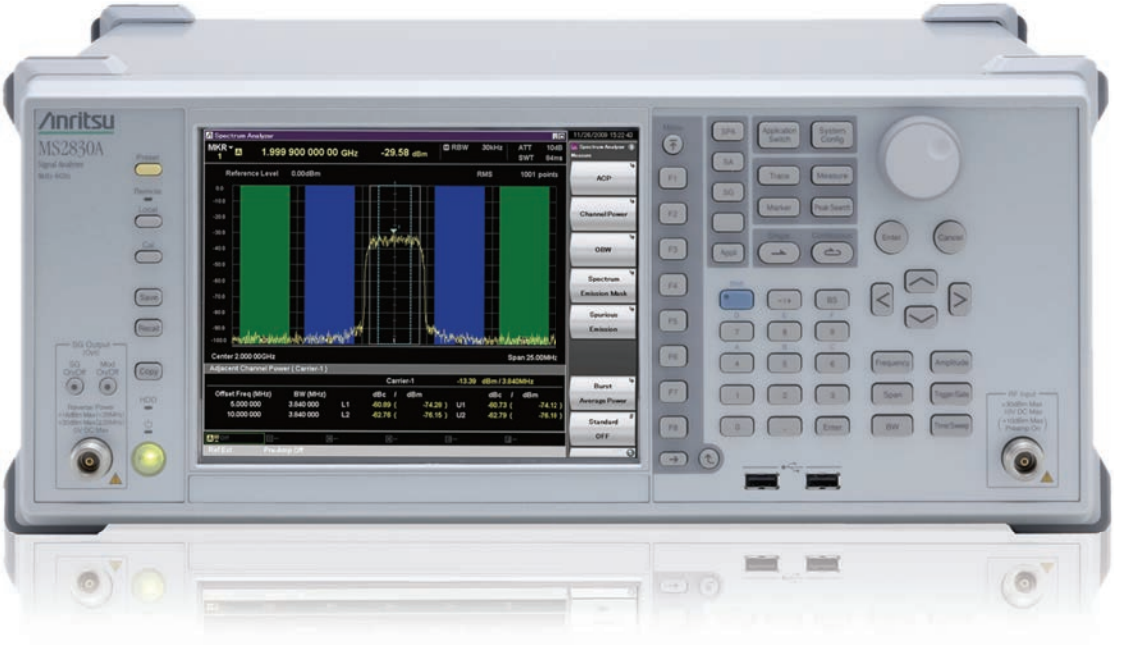
co-organizers, the WCSP conference series has established itself as the premier forum for the presentation of new advances and research results in the fields of wireless communications and signal processing.

WCSP 2014, the next edition of WCSP, will take place in Hefei, China, 23-25 Oct., 2014. The conference will be hosted by the University of Science and Technology of China. For more information about WCSP 2013 and WCSP 2014, please visit http://www.ic-wcsp.org.

Spectrum/Signal Analyzers

Since 1895

MS2830A



The New Standard

ONE Box, Many Uses, ONLY from Anritsu.

The MS2830A Signal Analyzer from Anritsu – smart, innovative design at its best. A Spectrum Analyzer, Signal Analyzer and Signal Generator, this single chassis has the unique ability to meet varying test needs efficiently.

FREE Guide to Spectrum Analysis www.goanritsu.com/2830IEEE11

1-800-ANRITSU
www.anritsu.com

Anritsu
envision : ensure

THE FUTURE OF WI-FI


Edward Au


Minho Cheong


Chiu Ngo


Carlos Cordeiro


Weihua Zhuang

Driven by the rapidly increasing demand for high data rate services and usage in a spectrum of application areas, wireless systems are compelled to evolve in order to meet the extraordinary performance requirements, especially in terms of spectral efficiency, coverage, latency, and energy efficiency. Regarding wireless local area networks, the first few widely accepted amendments to IEEE 802.11 wireless networking specifications, IEEE 802.11b/a/g, featured low spectral efficiencies, which are becoming insufficient to satisfy explosive traffic growth and the ever increasing consumer connectivity demand. With the soaring cost of the limited bandwidth at the 2.4 GHz frequency band, sustained improvement in spectral efficiency and quality of service has been achieved in IEEE 802.11n, which is mainly driven by advances in communication theory and the use of the 5 GHz frequency band. Specifically, the successive introduction of novel techniques such as multiple-input multiple-output (MIMO) antenna techniques and space-time coding, and the massive improvement in hardware and processing power, have enabled progressive system improvement. The introduction of advanced transmission techniques in recent years, notably multiuser MIMO and transmit beamforming, has provided additional powerful means for boosting the performance of Wi-Fi to gigabit per second speed and leading to the emerging IEEE 802.11ac. This is followed by the future introduction of IEEE 802.11ad products utilizing the 60 GHz frequency band, which are expected to take Wi-Fi speeds to multiple gigabits per second, and IEEE 802.1lax products, which are expected to improve the spectrum efficiency of Wi-Fi, thus enhancing system throughput per area in high-density scenarios of access points and client stations.

With seemingly limitless opportunities for new products and services, the proliferation of Wi-Fi continues to soar. Wi-Fi products can be deployed not only in apartments, but also in large corporations and campuses, and do everything from simple web browsing and peer-to-peer sharing to bandwidth-hungry and connectivity-demanding applications such as multimedia streaming and real-time teleconferencing, cable replacement, and wireless docking, to name a few. Coupled with the recent introduction of Wi-Fi CERTIFIED™ Passpoint, in which users can enjoy seamless and secure connectivity when roaming between cellular and Wi-Fi and between Wi-Fi networks, Wi-Fi devices now offer higher capacity and improved power management, and readily handle many demanding applications while paving the way for new products and services. Furthermore, the technology is now reaching beyond phones, PC networking, and consumer electronics into new sectors, such as the automotive industry and smart energy. According to Strategy Analytics 2012, Wi-Fi is now in over 25 percent of households, which is approximately 440 million people [1]. Furthermore, ABI Research recently increased its Wi-Fi shipment outlook and predicted that almost 3 billion Wi-Fi devices are expected to be shipped by 2015, which is nearly double the 1.5 billion devices shipped in 2012 [2].

The increasing popularity of Wi-Fi, combined with excellent market forecasts from reputable market research and intelligence firms, indicate that Wi-Fi is and will continue to be a key technology that is shaping the future of consumers and businesses worldwide. In response to this momentum of interest and popularity, our Feature Topic aims at providing a timely and concise reference to the state of the art, the latest research findings, and the future directions around Wi-Fi technologies.

An overwhelming number of papers were received for this Feature Topic, and five papers from a pool of high-quality submissions were selected based on their relevance to this Feature Topic and their technical merits. A number of good papers did not make the cut because of the above-mentioned criteria and the limitation of space. Nevertheless, we would like to thank all of the authors who submitted their work to this Feature Topic and all of our reviewers for their meticulous reviews, which were delivered in a timely fashion. In the following, we introduce the five articles by highlighting the contributions made therein. We hope our readers find these articles useful, not only in understanding the recent developments, but also for inspiring their own work.

Our Feature Topic begins with "Wi-Fi Could Be Much More" contributed by W. Sun *et al.* This article gives our readers an overview and high-level understanding of a number of recently released and ongoing specifications developed by IEEE 802.11 and the Wi-Fi Alliance. Furthermore, the authors have also outlined the most telling features and the corresponding advantages of these specifications, such as enhanced throughput, enlarged coverage, and ease of use.

Following a nice overview of these published and ongoing IEEE 802.11 specifications and the Wi-Fi Alliance CERTIFIED Passpoint program, the next two articles focus on architectural design and implementation challenges of WiGig technologies. First, Jo *et al.* present "Holistic Design Considerations for Environmentally Adaptive 60 GHz Beamforming Technology," which addresses the importance of considering an active adaptive beamforming algorithm that will be vital to accurately accommodate various surrounding environments for future 60 GHz applications. Then Rajagopal discusses an important topic that has yet to be well understood in the industry, power efficiency in future multiple gigabit Wi-Fi systems. In particular, the author presents a low-power architecture suitable for large-bandwidth Wi-Fi systems and discusses how the power efficiency challenge would be addressed.

The next two articles cover another very popular topic, the interworking of Wi-Fi and cellular systems. The first article is contributed by Kudo *et al.* and is titled "An Advanced Wi-Fi Data Service Platform Coupled with a Cellular Network for Future Wireless Access." Here, the authors present a Wi-Fi management architecture that utilize the potential capacity of Wi-Fi in the cellular network and provide high-grade user experience even in high-density Wi-Fi environments. In the second of these articles, "Enabling the Coexistence of LTE and Wi-Fi in Unlicensed Bands," Abinader *et al.* discuss performance issues that arise from concurrent operation of Wi-Fi and LTE in the same unlicensed bands from the radio resource management viewpoint. A few coexistence mechanisms and future research directions that may lead to a successful joint deployment of Wi-Fi and LTE are also presented.

REFERENCES

[1] http://www.strategyanalytics.com/default.aspx?mod=pressreleaseviewer&a0=5193.
[2] http://www.abiresearch.com/press/total-cumulative-wi-fi-enabled-device-shipments-re.

BIOGRAPHIES

EDWARD AU [SM] (edward.ks.au@gmail.com) is a senior staff member of Marvell Technology Group responsible for product certification and standardization of Wi-Fi and Bluetooth. He chairs a few technical task groups related to location, power saving, and smart grid technologies in the Wi-Fi Alliance and a study group on next generation 60 GHz in IEEE 802.11. He has a strong research record, having published tens of papers and patents. He also serves as Editor of various IEEE journals, and has served as a Track/Symposium Co-Chair of IEEE conferences. He is the receipt of the 2013 Top Editor Award of *IEEE Transactions on Vehicular Technology*.

MINHO CHEONG [SM] is a managing director at Newracom, Inc., ETRI's spin-off company which develops solutions for Korea Wi-Fi ecosystem. He has been a project leader, Special Fellow and head of delegates of IEEE 802 at ETRI, and worked on R&D on 4G systems, multi-gigabit-per-second nomadic system and next generation WLAN. He was the coordinator of Korea's standardization for next generation WLAN, Chair of VHT Working Group at TTA, and PHY Co-chair and Functional Requirements Editor of IEEE 802.11ac and IEEE 802.11ah. His research interests include OFDM, MIMO, and interference cancellation, on which he has filed over 100 patents. He and his group were named the "Nation-Wide Outstanding Research Group" by the Prime Minister of Korea in 2007. He was the recipient of the Silver Prize in Human-Tech. Thesis Contest in 2004 and the Grand Prize in the DSP Design Contest in 1997.

CHIU NGO [SM] is head of Standards and Technology Enabling, Samsung Electronics Silicon Valley R&D Center. As a senior director, he leads Samsung's U.S. standardization activities for consumer electronics. He received his Ph.D and B.Sc. (Honors) degrees in electrical engineering from the University of Southern California and the University of Hong Kong, respectively. He has actively participated in standardization organizations and holds Samsung's position on some Boards of Directors. He has co-authored more than 40 published papers and holds more than 150 U.S. patents. He is a Chartered Engineer of IET.

CARLOS CORDEIRO [SM] is a principal engineer in the Mobile and Communications Group within Intel Corporation. He is the overall lead of Intel's standardization programs in Wi-Fi and in the area of short-range multi-gigabits-per-second wireless systems using millimeter frequencies. In the Wi-Fi Alliance, he is a member of the Board of Directors and serves as the Technical Advisor, in addition to chairing the technical task group on 60 GHz. He was the Technical Editor of the IEEE 802.11ad standard. Due to his contributions to wireless communications, he received several awards including the prestigious Global Telecom Business 40 under 40 in 2012 and 2013, the IEEE Outstanding Engineer Award in 2011, and the IEEE New Face of Engineering Award in 2007. He is the co-author of two textbooks on wireless published in 2006 and 2011, has published about 100 papers in the wireless area alone, and holds over 30 patents. He has served as an Editor of various journals.

WEIHUA ZHUANG [F] has been with the University of Waterloo, Canada, since 1993, where she is a professor and a Tier I Canada Research Chair in Wireless Communication Networks. Her current research focuses on resource allocation and QoS provisioning in wireless networks, and on smart grid. She is a co-recipient of several best paper awards from IEEE conferences. She was the Editor-in-Chief of *IEEE Transactions on Vehicular Technology* (2007–2013), and Technical Program Symposia Chair of IEEE GLOBECOM 2011. She is a Fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada, and an elected member of the Board of Governors and VP Mobile Radio of the IEEE Vehicular Technology Society. She was an IEEE Communications Society Distinguished Lecturer from 2008 to 2011.

Wi-Fi Could Be Much More

Weiping Sun, Okhwan Lee, Yeonchul Shin, Seongwon Kim, Changmok Yang, Hyoil Kim, and Sunghyun Choi

ABSTRACT

Wi-Fi has become an essential wireless technology in our daily lives, although the original intention of its introduction was to replace Ethernet cable. In this article, we outline the most remarkable features introduced during its ongoing technological evolution in terms of three major directions: throughput enhancement, long-range extension, and greater ease of use. By stitching these advanced features together, we also envision a promising future that Wi-Fi technology will bring us in terms of spectrum heterogeneity, seamless service provisioning, and possible relations with cellular networks.

INTRODUCTION

Wi-Fi, the preferred term for IEEE 802.11 wireless local area networks (WLANs), has become an everyday tool for broadband Internet access in our daily lives.

Gradually, its position as the dominant carrier of wireless data traffic is being firmly cemented. According to the statistics in South Korea, the United States, Canada, Japan, Germany, and the United Kingdom, Wi-Fi contributed to about 73 percent of total wireless traffic on Android smartphones in April 2013, increased from 67 percent in August 2012 [1]. Such proliferation could be ascribed mainly to the support of the wide range of user devices (smartphones, tablet PCs, etc.), exploding network coverage, ongoing technological evolution, and the long-standing development of global standards.

This article offers a picture of paradigm shifts triggered by the development of Wi-Fi technologies currently underway. The picture, at its core, captures the idea that Wi-Fi, which was originally developed as an Ethernet cable replacement, has become an essential wireless technology in our daily lives, and will continue evolving to keep pace with spectrum availability and technological development.

The IEEE 802.11 Working Group (WG) released the first IEEE 802.11 standard, defining medium access control (MAC) and Physical (PHY) layers, in 1997, and has since adopted IEEE 802.11a, b, g, and n versions [2], with operations all restricted to the 2.4 GHz and 5 GHz unlicensed frequency bands. In its early stage, throughput enhancement was at the top of the list of the challenges faced by Wi-Fi. Starting with data rates up to 2 Mb/s (defined by the first standard), a number of significant advances have been made to enhance throughput.

The first step on the path to high-throughput WLAN was the introduction of orthogonal frequency-division multiplexing (OFDM) PHY, a popular technique that increases capacity by dividing a radio signal into multiple sub-signals which are transmitted simultaneously at different sub-carriers, first adopted by IEEE 802.11a in 1999. Although the maximum available data rates up to 54 Mb/s exposed Wi-Fi to more data-craving applications, the technology was far from a satisfactory solution until the advent of IEEE 802.11n in 2009.

The data rates defined in IEEE 802.11n are up to 600 Mb/s — more than 10 times 802.11a's 54 Mb/s. IEEE 802.11n was the first Wi-Fi standard with a speed comparable to that of wired networks (e.g., Ethernet). The key drivers of such significant improvements were the adoption of many cutting-edge technologies of the time, such as multiple-input multiple-output (MIMO), channel bonding, and frame aggregation, through which the efficiencies of spatial, spectral, and temporal resource utilizations were substantially enhanced. The triumph of IEEE 802.11n has resulted in unprecedented prosperity of Wi-Fi on both the technical and commercial fronts. "With 802.11n, there was a significant jump in minimizing the number of applications you had to keep a wired infrastructure for," said Dorothy Stanley, head of Standards Strategy at Aruba Networks.

In order to achieve prolonged growth, innovation, and vitality, Wi-Fi is expected to become more versatile and agile in dealing with its growing and diversified use in various scenarios such as indoor and outdoor, throughput and coverage, and personal and professional. Therefore, the IEEE 802.11 WG and Wi-Fi Alliance (WFA) continue to define and develop a number of advanced technologies that can be broadly classified into three broad categories: throughput enhancements, long-range extensions, and greater ease of use (Fig. 1).

Throughput Enhancements — From the beginning, high throughput has been a paramount concern for 802.11 WLAN. Several forces are still driving the trend of faster Wi-Fi technologies: the demand to extend its usability to more applications that otherwise required

Weiping Sun, Okhwan. Lee, Yeonchul Shin, Seongwon Kim, Changmok Yang, and Sunghyun Choi are with Seoul National University.

Hyoil Kim is with Ulsan National Institute of Science and Technology.

0163-6804/14/$25.00 © 2014 IEEE

wired infrastructure, and the need for more powerful wireless access technologies to support high-quality data-intensive applications, such as high-definition (HD) video streaming.

Long-Range Extensions — The current operating frequency bands, the 2.4 GHz and 5 GHz bands, have set limits on the transmission range of IEEE 802.11; hence, Wi-Fi has always been treated with indifference for outdoor environments. To make Wi-Fi more favorable to enlarged coverage, the Wi-Fi spectrum is being extended to other frequency bands.

Greater Ease of Use — As Wi-Fi functionality improves, its configuration and manipulation become more burdensome for users. The technology should be built on the premise of convenience.

In this article, we outline the most telling features in light of these three main directions that in which Wi-Fi technologies are advancing.

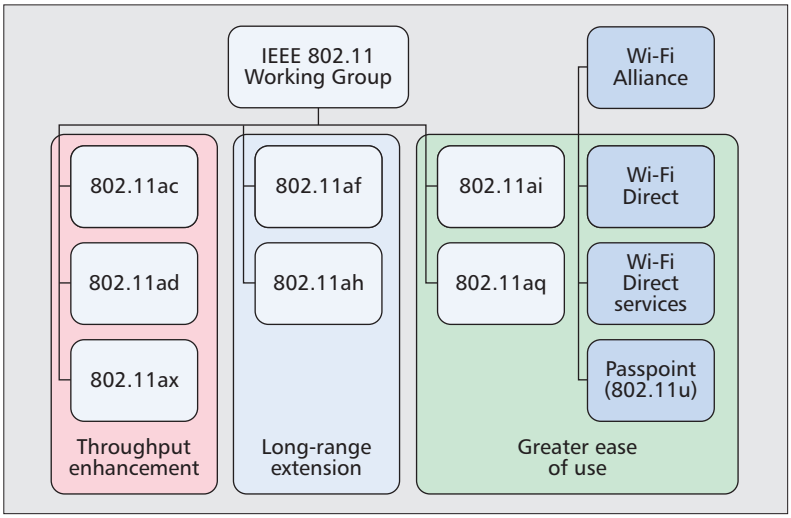


Figure 1. The evolution directions of Wi-Fi technologies.

THROUGHPUT ENHANCEMENTS

Two recently approved IEEE 802.11 amendments, IEEE 802.11ac [3] and IEEE 802.11ad [4], have been designed to follow the trend of faster Wi-Fi; the goal of both amendments is to provide theoretical maximum throughputs beyond 1 Gb/s [5]. "This level of performance has been a longtime goal of Wi-Fi proponents," stated Todd Antes, vice president of Product Management at Qualcomm Inc.

IEEE 802.11AC VERY HIGH THROUGHPUT

IEEE 802.11ac, a 5 GHz-only successor to 802.11n (Fig. 2a), improves the maximum throughput primarily by the following approaches: larger channel bandwidths of 80 and 160 MHz, multi-user MIMO (MU-MIMO), and higher-order modulation, that is, 256-quadrature amplitude modulation (QAM).

Wider Bandwidth Channels — The widening of channel bandwidth, called channel bonding, was first adopted in 802.11n, where the maximum channel bandwidth of 40 MHz is yielded by bonding two adjacent 20 MHz channels. When combining two channels, the theoretical data rate more than doubles since the guard band between the two bonded channels is removed.

IEEE 802.11ac takes further steps to support 80 MHz and optionally 160 MHz channels by bonding adjacent channels. Moreover, to increase the probability of composing a 160 MHz channel, 802.11ac also allows the generation of a 160 MHz channel by combining two physically non-adjacent 80 MHz channels, called 80+80 MHz.

Multi-User MIMO — Higher data rates can also be achieved with the multiple-antenna system known as MIMO. In the case of single-user MIMO (SU-MIMO), which is supported in 802.11n, the transmitted data is divided into multiple independent spatial streams and transmitted simultaneously via multiple antennas to a single receiver. MU-MIMO advances SU-MIMO by enabling an access point (AP) to transmit multiple spatial streams via multiple antennas to multiple receivers simultaneously.

MU-MIMO improves performance by serving multiple Wi-Fi clients in parallel rather than serially, as was the case in 802.11n, where the highest rate, 600 Mb/s, is available only when both AP and client are equipped with four antennas such that there are four spatial streams available to MIMO transmission. The number of antennas embedded in the client (e.g., smartphone or tablet PC), however, is usually limited to one or two due to the space limits of the device, although an AP with three to four antennas has become commonplace, resulting in the bottleneck of maximum data rate available in practice. MU-MIMO alleviates such inefficiency by enabling simultaneous reception at multiple clients so that the number of spatial streams is governed by the total number of antennas embedded in the clients, not per-client. IEEE 802.11ac supports downlink MU-MIMO only, with up to four receivers and up to eight spatial streams, thus doubling the number of supported spatial streams in 802.11n.

Higher-Order Modulation — The highest-order modulation in 802.11 WLAN has been 64-QAM ever since the adoption of 802.11a. IEEE 802.11ac newly adopts 256-QAM, thus enabling encoding four times as dense as the 64-QAM used by 802.11n.

In the 160 MHz mode (with 468 data subcarriers per OFDM symbol), a data rate of 866.7 Mb/s can be achieved with a single spatial stream using 256-QAM (i.e., 8 bits/sub-carrier/OFDM symbol), 5/6-rate coding, and a short guard interval: 8 (bits) × 468 (data sub-carriers) × (5/6) (code rate) × 277.8 (ksym/s). With the maximum number of spatial streams (eight), data rates up to 6.9 Gb/s are possible.

IEEE 802.11AD VERY HIGH THROUGHPUT

60 GHz Wi-Fi — IEEE 802.11ad, also known by its nickname "WiGig," defines the operation of WLAN over the unlicensed 60 GHz frequency

IEEE 802.11ad defines a fast session transfer between 802.11 PHY layers and the sustenance of the quality of experience (QoE) of existing 802.11 users. Therefore, a tri-band operation over the 2.4 GHz, 5 GHz, and 60 GHz bands with backward compatibility to the legacy 802.11 WLAN is newly defined.

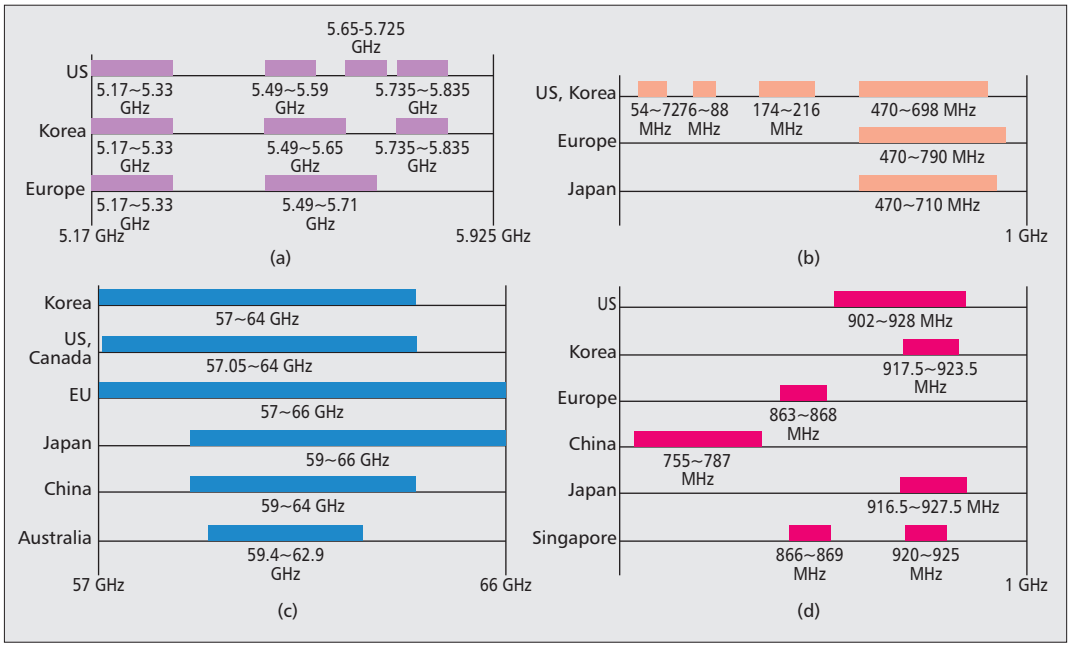


Figure 2. Frequency bands used by coming Wi-Fi technologies depending on regions: a) frequency bands of 802.11n/ac at 5 GHz; b) frequency bands of 802.11af at TVWS; c) frequency bands of 802.11ad at 60 GHz; d) frequency bands of 802.11ah below 1 GHz.

band, that is, the millimeter-wave (mmWave) band (Fig. 2c). Compared to 2.4 GHz and 5 GHz bands, communication over 60 GHz bands suffers from severe propagation loss and signal attenuation, thus resulting in a short communication range. On the other hand, it has an advantage of much broader available bandwidth. Moreover, thanks to the short wavelength in such a high-frequency band, a very large number of antennas can be deployed in a small area to form a high-directional beam, which concentrates the transmitted power to a particular direction and compensates for the signal attenuation. In this regard, 802.11ad is expected to be used for high-definition (HD) video transmission, high-rate data synchronization, and so on, while adaptive beamforming and multi-antenna configuration are becoming the core issues.

IEEE 802.11ad defines a fast session transfer between 802.11 PHY layers and the sustenance of the quality of experience (QoE) of existing 802.11 users. Therefore, a tri-band operation over the 2.4 GHz, 5 GHz, and 60 GHz bands with backward compatibility to the legacy 802.11 WLAN is newly defined.

PHY Feature — IEEE 802.11ad defines both single-carrier (SC) PHY supporting data rates up to 4620 Mb/s (with π/2-16QAM, 3/4-rate coding, and data symbol rate of 1540 Msym/s) and OFDM PHY supporting data rates up to 6756.75 Mb/s (with 64-QAM, 13/16-rate coding, 336 data sub-carriers, and OFDM symbol rate of 4125 ksym/s), both using 2.16-GHz-wide channels. The SC PHY is suitable for low-power mobile devices by virtue of its low power consumption. Besides, OFDM PHY can be adaptively used according to the link distance and the existence of obstacles for its longer communication range and greater resilience to delay spreads.

MAC Feature — The 802.11ad MAC, on the other hand, defines time-division multiple access (TDMA) above the existing contention-based carrier sense multiple access with collision avoidance (CSMA/CA) to support quality of service (QoS). It also supports high directivity with modifications on control frame operation such as beacon frames and clear-to-send (CTS) frames for direction-aware network allocation vector (NAV) allocation.

To be specific, a personal basic service set (PBSS) is defined in the 802.11ad MAC for peer-to-peer (P2P) communications. PBSS allows only a station chosen as a PBSS central point (PCP) to transmit beacon frames, possibly in different directions. Additional beamforming training and announcement after the beacon transmission allow directional MAC, thus enabling QoS guarantee and efficient power management.

IEEE 802.11ax High-Efficiency WLAN

Over the years, efforts on throughput enhancements (e.g., 802.11n/ac/ad) have been primarily focused on theoretical peak throughput in a single BSS environment. The tremendous progress made in this direction has brought us to a point where the emphasis has shifted to "real-world" performance.

Along with the growing population of Wi-Fi users, increasingly more APs are deployed in crowded areas to cater to both capacity and coverage demands. However, the goal is not likely to be achieved in reality simply by deploying more APs densely within a limited area; the resulting environments tend to be overlapping basic service sets (OBSSs), in which inter-BSS interference and collisions are likely to become more severe.

IEEE 802.11 TGax was recently established to address the challenges. IEEE 802.11ax, at its very early stage of standardization, aims to improve the efficiency of spectrum utilization by enhancing the area throughput (measured in bits per second per square meter) and average per-user throughput in both indoor and outdoor highly-dense deployment scenarios by advancing both PHY and MAC layers. It is the first time per-user throughput in multiple BSS environments are being considered, thus reflecting real-world performance more closely. Currently, TGax is considering state-of-the-art technologies including uplink MU-MIMO, orthogonal frequency-division multiple access (OFDMA), OBSS interference handling, and full duplex radio as 802.11ax key features.

Long-Range Extensions

Along with the great advances that 802.11ac and 802.11ad will bring in terms of speed, the 802.11 WG triggered two new standard extensions, IEEE 802.11af [6] and IEEE 802.11ah [7], for the purpose of long-range extensions at frequency bands below 1 GHz.

IEEE 802.11af TV White Space

TV white space (TVWS) is the temporarily vacant spectrum resources in very high frequency (VHF) and ultra high frequency (UHF) bands originally licensed to TV broadcasters and wireless microphones, which can be opportunistically utilized by unlicensed devices as long as no harmful interference is imposed on the licensed users. TVWS resides in 470–790 MHz in Europe and the United Kingdom, and non-continuous 54–698 MHz in Korea and the United States, as shown in Fig. 2b.

IEEE 802.11af defines WLAN operations at TVWS to deliver so-called Super Wi-Fi. Thanks to the favorable propagation characteristics of such low-frequency bands compared to 2.4 GHz and 5 GHz, including reduced path loss and better wall-penetrating ability, a Super Wi-Fi signal can travel longer distances than a typical Wi-Fi signal. Therefore, over-the-air broadband access can be implemented at lower cost by deploying 802.11af APs much less densely.

IEEE 802.11af mandates an operation under strict regulatory constraints, based on location-aware devices and online databases called geolocation databases (GDBs). A GDB stores location-specific information of available spectrum and usage schedule, and geolocation-capable 802.11af APs access the GDB via the Internet to obtain the necessary conditions to operate only where (geographically and spectrally) and when they do not interfere with nearby licensed devices in the TVWS.

802.11af is supposed to fulfill several requirements in terms of operating frequency spectra such as narrow channel bandwidth (6–8 MHz depending on the regions) and non-contiguous available channels due to the time-varying TVWS usage by TV users. Accordingly, it employs most advanced features of 802.11ac such as MU-MIMO by designing its PHY based on 40 MHz 802.11ac PHY, and supports both contiguous and non-contiguous channel bonding of up to four channels. For protection of TV users operating in adjacent channels, it also introduces additional guard bands, achieving 55 dB adjacent channel leakage ratio (ACLR).

IEEE 802.11ah below 1 GHz

Although 802.11af aims to provide a long-range Wi-Fi, the regulatory restrictions on the availability of spectral and temporal resources inherently limit its applicability in many locations, especially in urban areas, where many TV broadcast stations almost fully utilize TV bands already. Due to the intrinsic drawbacks of 802.11af and the increasing demand for ubiquitous wireless access, IEEE 802.11ah was initiated to specify the operation at unlicensed bands below 1 GHz (e.g., 917.5–923.5 MHz in Korea and 902–928 MHz in the United States), as shown in Fig. 2d.

IEEE 802.11ah is expected to provide a much improved transmission range compared to conventional Wi-Fi thanks to the superior propagation characteristics. Due to the long-range but limited bandwidth, 802.11ah is considered highly suitable for large-scale low-rate sensor networks (e.g., smart grid), where the number of involved devices in a given network could be much larger than that of conventional 802.11 Wi-Fi. On the other hand, target devices in the sensor networks are likely to be battery-powered; hence, the power saving features become critical to the performance of 802.11ah. Another challenge encountered by 802.11ah is the scarcity of available spectra, so increasing spectral efficiency is one of the main concerns in its protocol design.

In order to cope with such expected requirements, 802.11ah has introduced a number of enhancements in terms of power saving, the number of supported stations per AP (i.e., up to 8191 stations compared with 2007 stations of the legacy standard), medium access schemes (e.g., a new medium access scheme called restricted access window, RAW, has been proposed to mitigate collisions among a large number of stations by dividing time resource into several intervals, each of which is designated to a certain group of stations for channel access), and greater compactness of various frame formats [8]. Moreover, 802.11ah has designed a new PHY layer based on a 10 times down-clocked operation of 802.11ac PHY (making 802.11ah 10 times slower than 802.11ac), thus able to inherit 802.11ac PHY's advanced features.

Figure 3 illustrates the supported data rates and transmission ranges of the above-presented 802.11 standards. Table 1 also presents an overall performance comparison among them.

Greater Ease of Use

IEEE 802.11ai [9] and IEEE 802.11aq [10] aim to enhance user friendliness by reducing the initial link setup delay and providing pre-association service discovery, respectively. WFA also defines a number of new standards and certification programs, including Wi-Fi Direct [11], for direct communication among Wi-Fi devices without the aid of an AP, and Passpoint [12], for automatically joining a Wi-Fi subscriber service at hotspot areas.

Along with the great advances that 802.11ac and 802.11ad will bring in terms of speed, 802.11 WG triggered two new standard extensions, IEEE 802.11af [6] and IEEE 802.11ah [7], for the purpose of long-range extensions at frequency bands below 1 GHz.

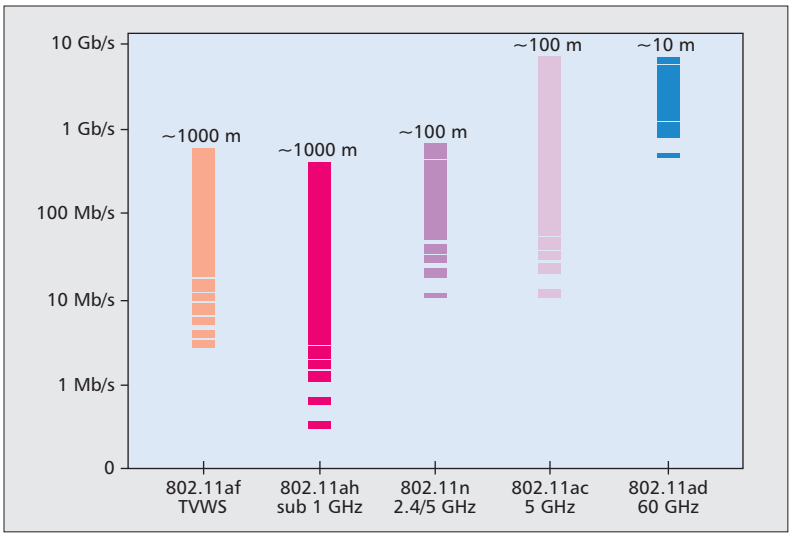


Figure 3. Supported data rates and transmission ranges of various 802.11 standards.

IEEE 802.11ai Fast Initial Link Setup

Typically, in order to use Wi-Fi service, a user should wait for a device to go through several steps before obtaining broadband Wi-Fi connectivity. The initial link setup — a technical term that specifies the procedures required for a first-time user to establish a secure Wi-Fi link with the most favorable AP — is, however, far from simple. The procedure basically consists of five steps: AP discovery, network discovery, authentication, association, and higher-layer configurations such as IP address configuration.

The challenges 802.11ai aims to address come from an environment where a large number of APs are densely deployed, and a massive amount of new users flock to the site. When these users simultaneously initiate link setup, the amount of traffic thus generated is likely to overwhelm the network capacity, and consequently, the time to wait for a connection setup exceeds the threshold users can tolerate. Therefore, there is a strong need for a more efficient and well scalable mechanism.

Accordingly, 802.11ai fast initial link setup (FILS) focuses on reducing the duration of the time spent in each step in order to complete the initial link setup within 100 ms. For example, the AP discovery time can be reduced by obtaining the information of the neighboring APs from another AP. Further optimization has also been made in both active and passive scanning. In active scanning, a station's probe request can be delayed or aborted by overhearing another station's probe request, and an AP's probe response can be broadcast instead of unicast so that all the nearby stations can acquire the AP information. The FILS Discovery (FD) frame, which conveys a part of the information of a beacon frame while being transmitted more frequently, is also designed to boost passive scanning performance.

IEEE 802.11aq Pre-Association Discovery

Wi-Fi is evolving into a more versatile technology that provides more than just Internet access. However, as service provisioning becomes more diverse, AP (or network) selection becomes more burdensome, still left to users' demand. This creates an opportunity for IEEE 802.11aq to help Wi-Fi users with the selection of the "right" AP by making more considerate information available to them before association.

For the delivery of service discovery information at the pre-association stage, technical modifications above the PHY layer are currently considered by TGaq. There are several existing higher-layer service discovery/description approaches, for example, Universal Plug and Play (UPnP), Bonjour, and Access Network Query Protocol (ANQP), as well as the mechanisms to deliver information at the pre-association stage, such as the IEEE 802.11u Generic Advertisement Service (GAS) framework; hence, TGaq will develop an approach by leveraging such existing schemes.

Wi-Fi Direct

"People tend to think of Wi-Fi as wireless Internet, but that's only one use of Wi-Fi," said Greg Ennis, WFA's technical director. Another step of Wi-Fi's evolution is to move into the P2P personal area networking realm, which until now has been the province of Bluetooth. This effort corresponds to the work being done in Wi-Fi Direct, the certification name of the Wi-Fi P2P specification defined by WFA to enable direct connections among devices without the help of an AP [11].

In order to inherit the advantageous features of traditional Wi-Fi (e.g., power saving for stations), Wi-Fi Direct mimics the infrastructure-based WLAN architecture. That is, Wi-Fi Direct devices form a group called a P2P group, where a group member, called the group owner (GO), works like an AP in the infrastructure-based WLAN. From the users' perspective, these devices provide P2P communication in the sense that the GO's identity is not revealed to the users while the GO is dynamically selected during the group formation stage.

To construct a P2P group, two devices should find each other first via the find phase operation, which is done by conducting active scanning at three "social" channels at 2.4 GHz (i.e., channels 1, 6, and 11). Then several subsequent steps, such as GO negotiation, Wi-Fi protected setup (WPS) provisioning, and IP address configuration, are taken. By completing all these steps, a device becomes a GO and serves other P2P clients via a secure wireless link.

Besides the power saving feature inherited from 802.11 WLAN, which is dedicated for stations, Wi-Fi Direct defines two novel power saving mechanisms for the GO, opportunistic power saving and notice of absence (NoA), since the GO is also likely to be a normal battery-powered portable device. Opportunistic power saving offers the GO a series of intermittent power saving opportunities by exploiting the time when every associated P2P client is in the doze state. NoA, by contrast, defines more active power saving operations that allow the GO to be absent for a scheduled duration by reporting its absence to the associated P2P clients in advance.

By utilizing Wi-Fi Direct, a mirroring service named Miracast has been standardized by WFA, and it allows multimedia contents, such as audio

	802.11ac	802.11ad	802.11af	802.11ah	802.11ax (expected)
Freq. spectrum	5 GHz	60 GHz	54–790 MHz	<1 GHz	2.4 & 5 GHz
Nominal range	~100 m	~10 m	~1 km	~1 km	~100 m
Channel bandwidths	20/40/80/160/ 80+80 MHz	2.16 GHz	6/7/8/12/14/16/ 24/28/32/ 6+6/7+7/8+8/ 12+12/14+14/16+16 MHz	1/2/4/8/16 MHz	—
Max data rate	6.933 Gb/s	6.756 Gb/s	568.9 Mb/s	346.7 Mb/s	—
Max mandatory rate	292.5 Mb/s	2.08 Gb/s	26.7 Mb/s	6.5 Mb/s	—
Key features	Downlink MU-MIMO, channel bonding, higher-order modulation	Beamforming	GDB-based channel access	Deep power saving, increased number of supported stations per AP, short frames, efficient medium access schemes	Uplink MU-MIMO full duplex, OFDMA, OBSS handling

Table 1. Comparison among upcoming Wi-Fi technologies.

and video, to be shared across devices seamlessly via Wi-Fi Direct connection. That is, it enables users to mirror the screen of a portable device (e.g., smartphone) onto a large screen TV or monitor in order to enjoy the entertainment more comfortably.

Even so, the lack of upper-layer applications has been one of the major handicaps that impede the debut of Wi-Fi Direct as a mainstream P2P technology. Correspondingly, a framework called Wi-Fi Direct services (WFDS) is under development to provide third-party developers a normalized platform interface so that the extensibility and interoperability of the resulting applications can be obtained easily, ultimately encouraging Wi-Fi Direct to become more essential in the future.

Passpoint

A new certification program called Passpoint has also been developed by WFA as an industry-wide solution to streamline the network access in hotspot areas [12]. Based on IEEE 802.11u [2] and WFA Hotspot 2.0 specifications, it eliminates the need for users to search and choose a network, to request the connection to the AP, and, in many cases, to re-enter their authentication credentials each time they initiate a Wi-Fi connection. Passpoint automates the entire process by enabling a seamless connection between hotspot networks and mobile devices while delivering a secure wireless link.

Envisioning the Future of Wi-Fi

In this section, we attempt to envision the future direction of Wi-Fi evolution by taking into account the aforementioned trend in Wi-Fi development.

More Versatile Wi-Fi Exploiting Spectrum Heterogeneity

Traditionally, the use of higher frequency bands has been a driving force for providing higher-speed Wi-Fi, mainly due to the availability of broader bandwidth. Nevertheless, the inferior propagation characteristics of such higher frequency bands have resulted in less competence in network coverage. The trade-off between capacity and coverage is well demonstrated in Fig. 4, which exhibits the maximum supported data rates according to the communication range regarding various Wi-Fi standards operating in 2.4 GHz, 5 GHz, and TVWS.

Diversification of Wi-Fi spectrum could introduce a more versatile Wi-Fi by adaptively integrating heterogeneous Wi-Fi standards according to the user context and network conditions [13]. For instance, an AP that jointly supports 2.4 GHz, 5 GHz, and TVWS may also be able to achieve the network performance corresponding to the upper envelope of the combined plots in Fig. 4 with multi-band multi-standard Wi-Fi users.

All-Wi-Fi Seamless Service Provisioning

Technological evolution and diversified frequency spectra have made Wi-Fi powerful and agile enough to be suitable for various environments, not solely restricted to indoor use. Combining the advantages of enhanced throughput, enlarged coverage, and easier use, we can envision ubiquitous broadband wireless access provided by a "Wi-Fi ecosystem," in which various Wi-Fi technologies take complementary roles for seamless service provisioning, as illustrated in Fig. 5.

For better understanding of the Wi-Fi ecosystem, here is a scenario that might happen in the near future:

Bob, a salesman, is watching TV at home in the morning. The HD contents are being transferred from a set-top box to the TV via 802.11ad, which replaces the traditional complex cable connections, giving the layout of the TV a higher degree of "freedom." At the same time, his tablet PC is connected to an 802.11ac AP, through which

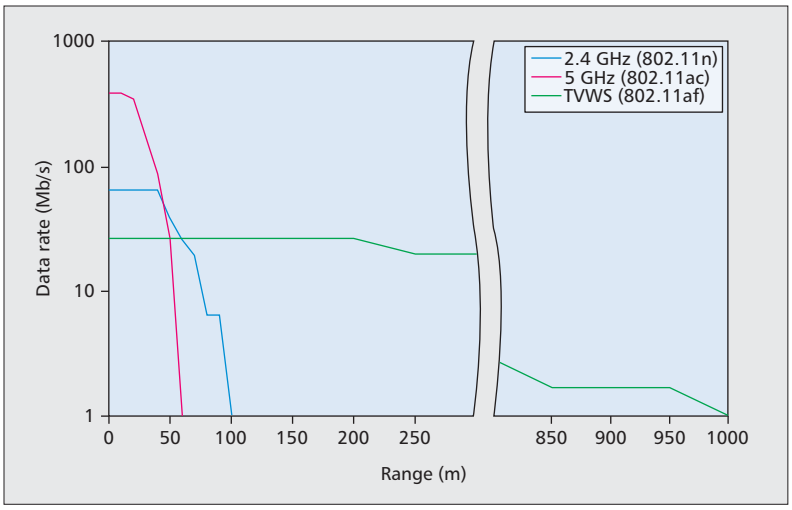


Figure 4. Trade-off between capacity and coverage among 2.4 GHz, 5 GHz, and TVWS Wi-Fi.

he enjoys smooth, flawless, and real-time video chat with his colleague, discussing today's meeting agenda. On his way to the subway, his phone tells him that there is an email to which he immediately responds using an outdoor Wi-Fi wirelessly backboned by 802.11af. At the subway station, his phone promptly discovers and associates with the "best" AP with the help of 802.11ai's FILS and 802.11aq's pre-association service discovery. During the commute, he watches video clips streamed from remote cloud servers via 802.11ax, which operates at high speed even in the highly dense environment with hundreds of other commuters, thanks to its high-efficiency design.

Although several technical challenges to make the scenario happen remain, Wi-Fi technology is evolving fast, filling the gap between reality and imagination.

Relationship with Cellular

Long-Term-Evolution (LTE), one of the most prominent cellular deployments across the world, is currently operating in the licensed spectrum (e.g., from 700 MHz to 2.6 GHz) to provide services in a more controlled manner than Wi-Fi operating in the unlicensed spectrum, by virtue of the exclusive spectrum occupancy. Recently, however, with the pressing need for additional spectral resources incurred by ever-increasing mobile traffic demand, the idea of deploying an LTE system in unlicensed bands (particularly the 5 GHz unlicensed band mostly used by Wi-Fi today) is on the horizon.

On the other hand, Wi-Fi, which thus far has been used by operators as a secondary carrier in indoor environments for the purpose of cellular traffic offloading, attempts to extend its territory to outdoor environments by increasing spectrum heterogeneity via the development of both 802.11af and 802.11ah.

As we have witnessed, the gap in spectrum usage policy, environments, and performance between these two major wireless systems, Wi-Fi and cellular, is narrowing along with the technological evolution [14]. Although we cannot yet anticipate what will happen to these two compelling options for mobile users, the possible anticipated outcomes include coexisting in a common system and becoming more tightly integrated, one of them being eliminated in a fierce competition, or a completely new wireless mobile system emerging, replacing both. Hence, it will be quite interesting to watch these innovative technologies unfold before us in the future.

Concluding Remarks

While Wi-Fi has become a dominant carrier of wireless data traffic, it continues evolving to keep pace with spectrum availability and technological development. In this article, we outline the most telling features of the Wi-Fi technologies being developed by the IEEE 802.11 WG and WFA in terms of their advantages of enhanced throughput, enlarged coverage, and easier use.

Several paradigm shifts that will probably occur in the near future have also been envisioned. First, diversification of Wi-Fi spectrum can improve Wi-Fi's versatility so that both coverage and capacity can be achieved by adaptively exploiting its augmented heterogeneity. Second, an all-Wi-Fi ecosystem is likely to be constructed by combining all the superior features of Wi-Fi technologies, in which seamless service provisioning can be provided with good service quality. Last but not least, Wi-Fi's relationship with the cellular network has so far been complementary, which might undergo revolutionary changes along with the ongoing evolution of Wi-Fi.

References

[1] M. Roberts, "Understanding Today's Smartphone User: An Updated and Expanded Analysis of Data-Usage Patterns in Six or the World's Most Advanced 4G LTE Markets," White Paper, June 2013.
[2] IEEE 802.11-2012, Part 11, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," Mar. 2012.
[3] IEEE 802.11ac, Part 11, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Enhancements for Very High Throughput for Operation in Bands Below 6 GHz," Dec. 2013.
[4] IEEE 802.11ad, Part 11, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Enhancements for Very High Throughput in the 60 GHz Band," Dec. 2012.
[5] L. Verma, M. Fakharzadeh, and S. Choi, "Wi-Fi on Steroids: 802.11 AC and 802.11 AD," *IEEE Wireless Commun.*, vol. 20, no. 6, 2013, pp. 30–35.
[6] IEEE 802.11af, Part 11, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Television White Spaces (TVWS) Operation," Dec. 2013.
[7] IEEE P802.11ah/D1.2, Part 11, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Sub 1 GHz License Exempt Operation," Feb. 2014.
[8] W. Sun, M. Choi, and S. Choi, "IEEE 802.11 ah: A Long Range 802.11 WLAN at Sub 1 GHz," *J. ICT Standardization*, vol. 1, no. 1, 2013, pp. 83–108.
[9] IEEE P802.11ai/D1.3, Part 11, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Fast Initial Link Setup," Feb. 2014.
[10] Y. Yang and D. Gal, "Proposed Specification Framework for TGaq," IEEE 802.11-13/0300r1, Mar. 2013.
[11] Wi-Fi Peer-to-Peer (P2P) Tech. Spec., v. 1.1, Wi-Fi Alliance, P2P Tech. Group, Oct. 2010.
[12] "Wi-Fi Certified Passpoint," http://www.wi-fi.org/discover-wi-fi/wi-fi-certified-passpoint.
[13] H. Kim, K. Shin, and C. Joo, "Downlink Capacity of Super Wi-Fi Coexisting with Conventional Wi-Fi," *Proc. IEEE GLOBECOM*, 2013.
[14] Huawei, "LTE Small Cell vs. Wi-Fi User Experience," White Paper, 2013.

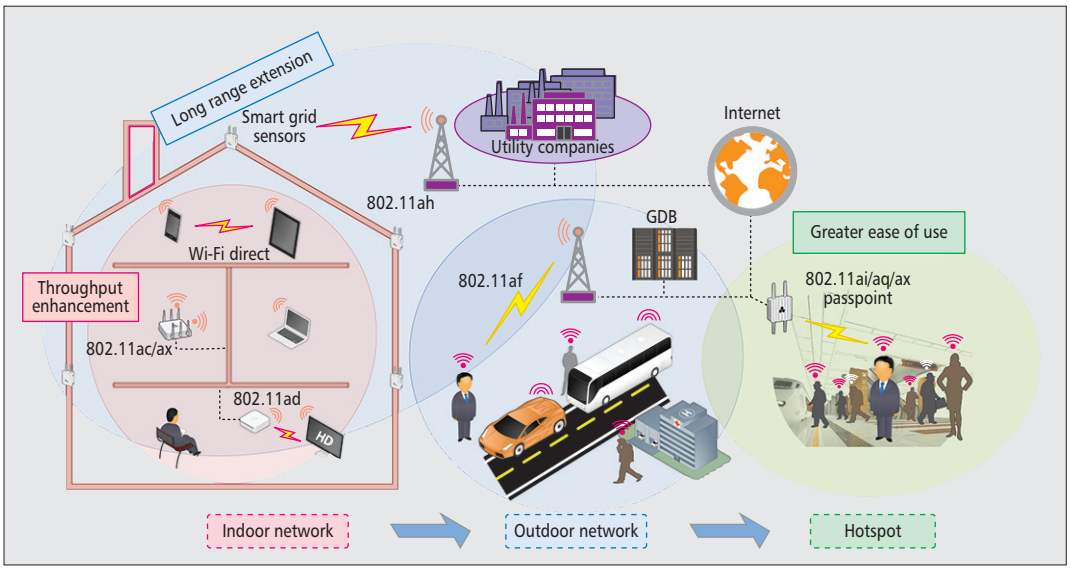


Figure 5. New paradigm of all-Wi-Fi heterogeneous access.

Combining the advantages of enhanced throughput, enlarged coverage, and easier use, we can envision ubiquitous broadband wireless access provided by a "Wi-Fi ecosystem," in which various Wi-Fi technologies take complementary roles for seamless service provisioning.

Biographies

Weiping Sun [S'12] received a B.E. degree in network engineering from Dalian University of Technology, China, in 2010. He is currently working toward a Ph.D. degree in the Department of Electrical and Computer Engineering, Seoul National University, Korea. His current research interests focus on protocol design and performance evaluation of Wi-Fi networks.

Okhwan Lee [S'09] received a B.S. from the School of Electronics Engineering at Seoul National University in 2006. He is currently pursuing a Ph.D. degree in the Department of Electrical Engineering and Computer Science, Seoul National University. His current research interests are in the area of wireless/mobile networks with emphasis on IEEE 802.11-based wireless mesh networking and the design of enhanced MAC/PHY protocols.

Yeonchul Shin [S'12] received his B.S. degree from the Department of Electrical Engineering of Seoul National University in 2010. He is currently working toward a Ph.D. degree in the Department of Electrical Engineering and Computer Science, Seoul National University. His current research interests include video multicast, QoS support, power management, algorithmic design, and performance evaluation for wireless networks, in particular, IEEE 802.11 WLANs.

Seongwon Kim [S'12] is a Ph.D. student at the Department of Electrical and Computer Engineering, Seoul National University. He received his B.S. degree in electrical engineering from Pohang University of Science and Technology in 2011, and his M.S. degree from Seoul National University in 2013. His current research interests include IEEE 802.11 WLAN, enterprise networks, and next-generation mobile networks.

Changmok Yang [S'14] received his B.S. degree in the Department of Electrical Engineering at Seoul National University in 2013. He is currently working toward a Ph.D. degree in the Department of Electrical Engineering and Computer Science, Seoul National University. His current research interests include protocol design and performance evaluation for wireless network, in particular, IEEE 802.11 WLANs.

Hyoil Kim [M'10] is an assistant professor at the School of ECE, UNIST, Korea, since 2011. Before joining UNIST, he was a postdoctoral researcher at the IBM T. J.Watson Research Center, 2010-2011. He received his B.S. degree in electrical engineering from Seoul National University in 1999, and M.S. and Ph.D. degrees in electrical engineering: systems from the University of Michigan in 2005 and 2010. His research interests include cognitive radios, mobile cloud computing, and next-generation WLAN. He served as a TPC member of IEEE GLOBECOM (2011-2014), ICUFN (2012-2015), and IEEE INFOCOM 2014, and as a publicity co-chair of ACM WiNTECH 2013.

Sunghyun Choi [S'96, M'00, SM'05, F'14] is a professor in the Department of Electrical and Computer Engineering, Seoul National University. Before joining the university in 2002, he was with Philips Research USA. He received his B.S. (summa cum laude) and M.S. degrees from the Korea Advanced Institute of Science and Technology, and received a Ph.D. from the University of Michigan. He has co-authored over 180 papers and holds about 120 patents. He is serving as an Editor of *IEEE Wireless Communications*. He was named an IEEE Fellow in 2014 for his contribution to the development of WLAN protocols.

Holistic Design Considerations for Environmentally Adaptive 60 GHz Beamforming Technology

Ohyun Jo, Wonbin Hong, Sung Tae Choi, SangHyun Chang, ChangYeul Kweon, Jisung Oh, and Kyungwhoon Cheun

ABSTRACT

Proliferation of all forms of smart devices has driven the demand for high-data-rate Wi-Fi in recent years. Utilizing wide bandwidth in high frequency is the most efficient way of guaranteeing the explosive data demands. In spite of the notable advancements of RF technologies at millimeter wave, the relatively high propagation loss remains problematic. The authors present efficient beamforming technology to mitigate this challenge. In this article, we first discuss the principles and major applications of beamforming technology. Afterward, we present the challenges for optimizing the efficiency of beamforming as well as the solutions to tackle them by proposing enhanced algorithms and advanced design architectures based on our hands-on experience and knowledge. Lastly, experimental results are discussed to deduce insight and vision of future Wi-Fi.

INTRODUCTION

The IEEE 802.11 family of standards represent wireless local area network (WLAN). Amendments for backward compatible interoperability and expeditious establishments of newer standards supporting enhanced data rates are the pillars of the evolution of Wi-Fi [1, 2]. The IEEE 802.11ad is one classical example. The standard was originally introduced as wireless gigabit (WiGig) in May 2009 by the Wireless Gigabit Alliance (WGA), which was an industry association including Intel, Samsung, Dell, Broadcom, and Marvell with the mission of developing and promoting the adoption of multi-gigabit-speed wireless communications technologies operating over the unlicensed 60 GHz band.

The medium access control (MAC) and physical (PHY) layers undergo a drastic makeover for the IEEE 802.11ad standard. The MAC includes new network architectures and functions: superframe structure, beamforming, multi-band operation, and and so. The PHY supports four physical modes: control PHY mode for beaconing and initial network entry, single carrier (SC) PHY mode for handheld devices, orthogonal frequency-division multiplexing (OFDM) PHY for high-performance applications, and low-power SC PHY for power-limited devices. These modifications collectively enable the wireless data throughput to reach up to 6.7 Gb/s, and can be applicable for uncompressed high definition (HD) video streaming service and high-rate data service [3].

In order to realize extremely high throughput, the IEEE 802.11ad standard utilizes 2.16 GHz bandwidth. However, in comparison to legacy bands, 60 GHz signal incurs much longer signal attenuation during propagation through the wireless air interface and the large amount of oxygen absorption further exacerbates this phenomenon [4–6]. Consequently, several key technologies have intensively been studied to overcome this inevitable problem [7]. As one of the most effective solutions to the huge propagation loss, beamforming plays a key role by preventing the transmission energy from spreading in every direction. Using an array antenna, electrical energy can be concentrated on the intended destination point.

In this article, we describe all the layers composing IEEE 802.11ad down to the practical details including its implementation and experimental results. We first start with comprehensible physics and signal processing to convey the principles of beamforming technology. We then introduce beamforming protocols defined in the IEEE 802.11ad standard and relate how the beamforming technology can be applied to future Wi-Fi scenarios. In addition, novel algorithms for enhancing the quality of service (QoS) from the perspective of the user are extracted and summarized to deduce insights on effective management and use of beamforming technology. The next section presents an in-depth description of the practical design and implementation of hardware-aided beamforming carried out by the authors at Samsung Electronics. The major blocks consisting of digital integrated circuits (ICs), RFICs, and antenna experience full-scale improvement and is evaluated at the system level through extensive experiments. The final section concludes the article.

The authors are with Samsung Electronics Co., Ltd., Suwon, Republic of Korea.

0163-6804/14/$25.00 © 2014 IEEE

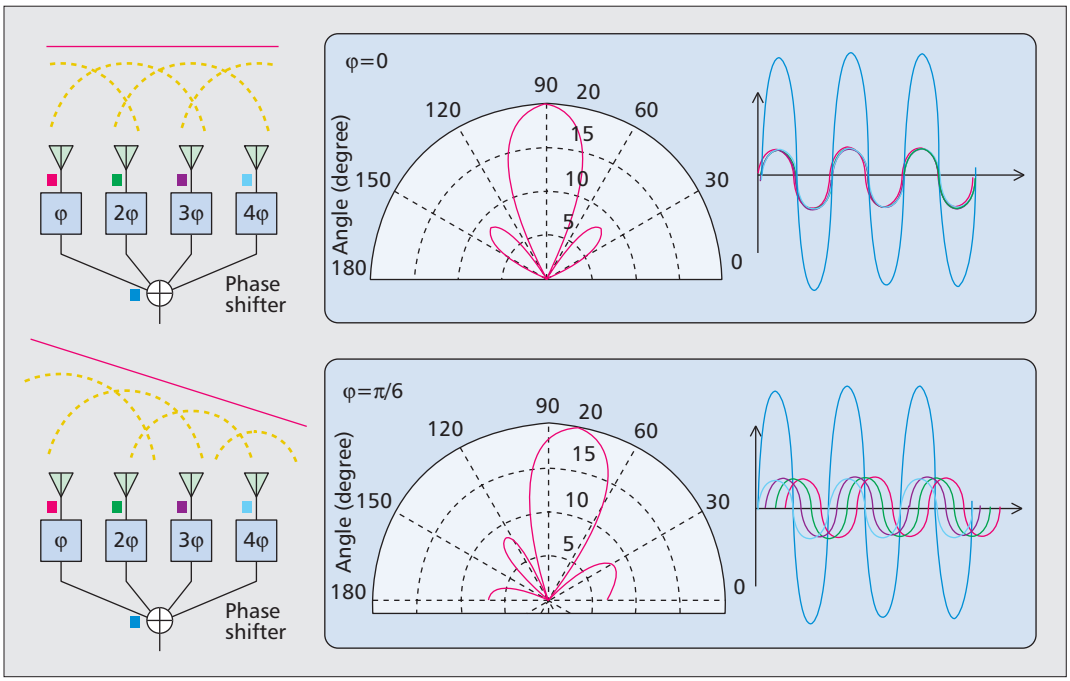


Figure 1. The principle of constructive/destructive superposition for beam steering.

All frame formats used in the beamforming protocols are defined in accordance with the frame structure of the existing Wi-Fi. Likewise channel access method for obtaining transmit opportunity(TXOP) originates from the existing Wi-Fi.

The Principle of mmWave Beamforming

The IEEE 802.11ad system exploits a millimeter-wave (mmWave) array antenna beamforming technology in the unlicensed 57–66 GHz frequency band (conventionally, 60 GHz band). The beamforming process in IEEE 802.11ad utilizes an electronically controlled beam-steering technique, so the directionality of the array beam pattern can be tuned to a desired direction (e.g., the direction of the propagation path that provides the least transmitted power attenuation), as shown in Fig. 1. The IEEE 802.11ad system may enable an mmWave phased array transmit and/or receive antenna, each antenna element of which is combined with a phase rotator [8].

The physical principle of the phased array antenna can be described as follows. When the incident wave impinges on the phased array receive antenna, the time of arrival at each array antenna element may be different. Note that the time-of-arrival difference depends on the spacing between antenna elements with respect to the impinging direction. To superimpose the received signals in a constructive fashion (i.e., to achieve high antenna gain effectively), the phased array receiver can compensate for the phase difference and combine the coherent signals. Due to the reciprocity, the operation of the transmit array antenna is similar to that of the receive array antenna. The array antenna pattern depends on the number of antenna elements, the spatial configurations of the antenna elements, and the gain pattern of the antenna elements. For example, we can manipulate the antenna pattern to be narrower and higher gain by using more antenna elements.

Beamforming Protocols of IEEE 802.11ad

The IEEE 802.11ad standard mainly describes three beamforming protocols as mandatory beam training mechanisms[9]: sector-level sweep (SLS), beam refinement protocol (BRP), and beam tracking. Figure 2 illustrates each of aforementioned beamforming protocols. All frame formats used in the beamforming protocols are defined in accordance with the frame structure of the existing Wi-Fi. Likewise, the channel access method for obtaining transmit opportunity (TXOP) originates from the existing Wi-Fi.

Sector-Level Sweep

SLS is devised for initial link detection by selectively training transmit antenna or receive antenna. Figure 2 shows the transmit sector sweep (TXSS) case for training a transmit antenna. During TXSS, several sector sweep (SSW) frames containing the same contents are transmitted to the peer station while continuously changing direction. Meanwhile, the receiver retains its receive antenna configuration as quasi-omni mode. It is also possible to carry out a receive sector sweep (RXSS), in which several SSW frames are transmitted in a quasi-omni pattern while the receiver continuously changes the receive antenna configuration. A station that has successfully received SSW frames notifies the best sector to the peer station through the response frame, which contains the best sector number. This information can be included in SSW frame and sector sweep feedback (SSFB) frame as shown in Fig. 2.

Beam Refinement Protocol

During the BRP process, a station improves its antenna configuration by using an iterative procedure. Basically, the transmit training/

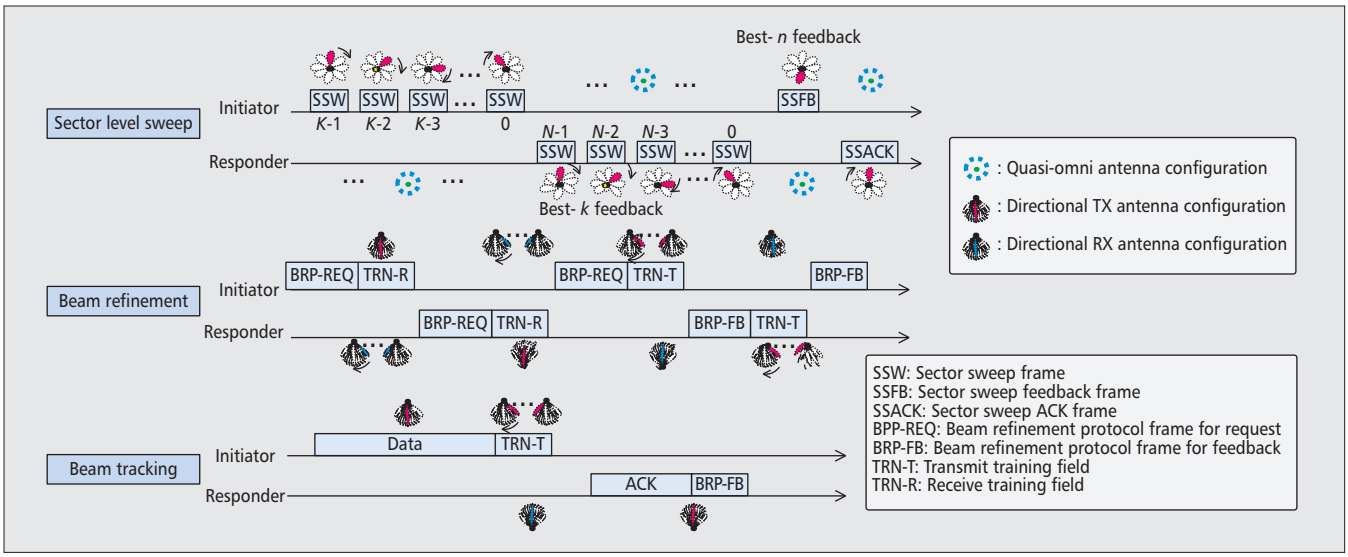


Figure 2. Beamforming protocols of IEEE 802.11ad.

receive training (TRN-T/R) field, which consists of Golay sequences, enables transmit/receive antenna training. When the TRN-T field is being transmitted, the transmitter continuously changes the transmit antenna configuration for transmit training, and the receiver retains its receive antenna configuration. Likewise, the TRN-R field is transmitted for training of the receive antenna configuration. When the TRN-R field is being transmitted, the receiver continuously changes its receive antenna configuration and the transmitter retains its transmit antenna configuration. The BRP frame used in the iterative process for beam training contains the training request and feedback information. The TRN-T/R field is appended to the BRP frame in case the training procedure is requested by the peer station.

Beam Tracking

Beam tracking is differentiated from the aforementioned beamforming protocols by the frame type used during beam training. The training field is appended to data frames in order for the training process to be carried out without discontinuity of data transmission. In this case, the beamforming request/feedback information cannot be obtained from the payload in the frame. The beamforming request information is conveyed in the physical layer convergence procedure (PLCP) header which contains information of the PLCP service data unit (PSDU). The beamforming feedback information is delivered in the next BRP frame. Synthetically, beam tracking is limited in iterative training, but still advantageous for its small signaling overhead and possibility of seamless data transmission.

Beamforming Algorithms for QoS Enhancement

In this section, we introduce enhanced algorithms that can explicitly improve the performance of beamforming in terms of QoS. The beamforming procedure for searching a new communication path should be carried out immediately and accurately when the link quality is degraded. Also, the training results should reflect the user's service requirements. Related to this fact, we propose cross-layer algorithms that can be generally applicable.

Consideration of Effective Metrics across Multiple Layers

Channel quality information, such as signal-to-noise ratio (SNR) and received signal strength indicator (RSSI), is the most typical indicator to describe link conditions. In most cases, channel quality information is obtained from the physical layer, and normally applied as a metric for beam training. Despite the fact that the channel quality information may possibly reflect the condition of a wireless link, the upper layer information is more relevant to the performance from the end users' perspective. However, the upper layer information is relatively more difficult to access than the physical information. Therefore, various information from multiple layers is collected within the allowable overhead range, as illustrated in Fig. 3a. An efficient architecture in which the information of the multiple layers can be tightly managed is proposed in the following subsection.

Efficient Protocol and Parameter Decision

The choice of protocol can have a great effect on the performance. For example, SLS is efficient for initially searching wide coverage by using coarse beams when the link is severely corrupted. On the other hand, it requires relatively long training time. Other beamforming protocols using training fields, such as TRN-T/R, are more suitable for fine training by generating less overhead. Likewise, they are inappropriate scenarios where the link condition is severely changed (e.g., the current wireless path is totally blocked). Therefore, the appropriate beamforming protocol should be selected while considering the

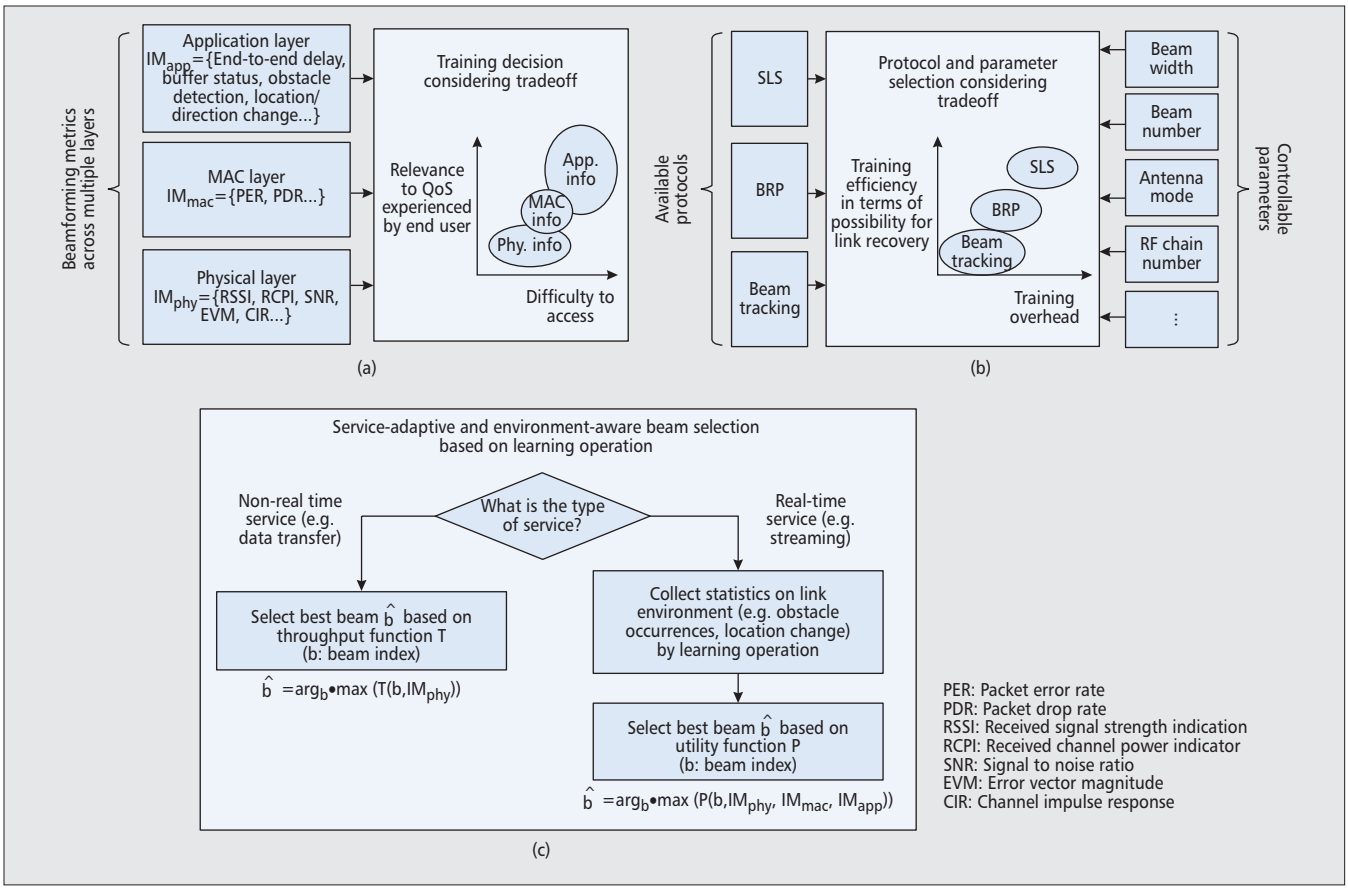


Figure 3. Beamforming algorithms for QoS enhancement: a) link quality evaluation and beamforming triggering by using cross-layer information; b) optimized beamforming protocol and parameter selection for fast and accurate link recorvery; c) adaptive beam selection by using awareness of the link environment obtained from upper layers.

cross-layer link information in Fig. 3a. Fine-tuned beamforming parameters to the link environment are also helpful in optimizing the beamforming performance. This indicates that the beamforming process can be situationally accommodated to the given condition by controlling parameters, such as beam width, number of beams used in the device, and the configurations of RF and antenna. In Fig. 3b, we applied the trade-off between the training efficiency and training overhead to perform the beamforming protocols effectively by optimizing the beamforming parameters.

Service-Adaptive Beam Selection with Environmental Awareness

The standard generally does not describe specific beam selection criteria. However, selecting the best beam situationally may significantly impact the user experience since the best beam can be defined differently according to the service. The user scenario should be reflected in the selection criteria. For example, the best beam for data transfer can be the one with the maximum throughput. On the other hand, for a real-time service such as video streaming and interactive games, stability of the service is the most important criterion. The most robust beam with minimal fluctuation and discontinuity of service is the best beam in this case. Thus, we can differentiate the beam selection algorithm and the figures of merit depending on the requirements of the service.

For realizing this, we added a learning operation to collect the beam statistics on the link environments. Physical layer information as well as historical and environmental information are gathered and maintained for each beam from upper layers. The collected information includes the external effects on the communication link such as blocking frequency due to obstacles and the existence of interferers. Based on this cross-layer information, the priority is evaluated to select the best beam taking into account the environmental awareness and beam characteristics as shown in Fig. 3c. For non-real-time services, the priority function can be replaced to the simple throughput function solely using physical channel information.

Advanced Design and Implementation of Hardware-Aided Beamforming

Significant advancements at the hardware level are imperative to support the proposed service-adaptive beam selection methodology presented above. Novel improvements range across the digital, RF, and antenna layers, and the pro-

In the proposed design, essential functionalities are included in the shape of digital circuits for boosting the beamforming process and for obtaining the advantages of HW implementation. Handling beamforming process in HW is very challenging as expected. However it returns great benefits when the hurdles are overcome.

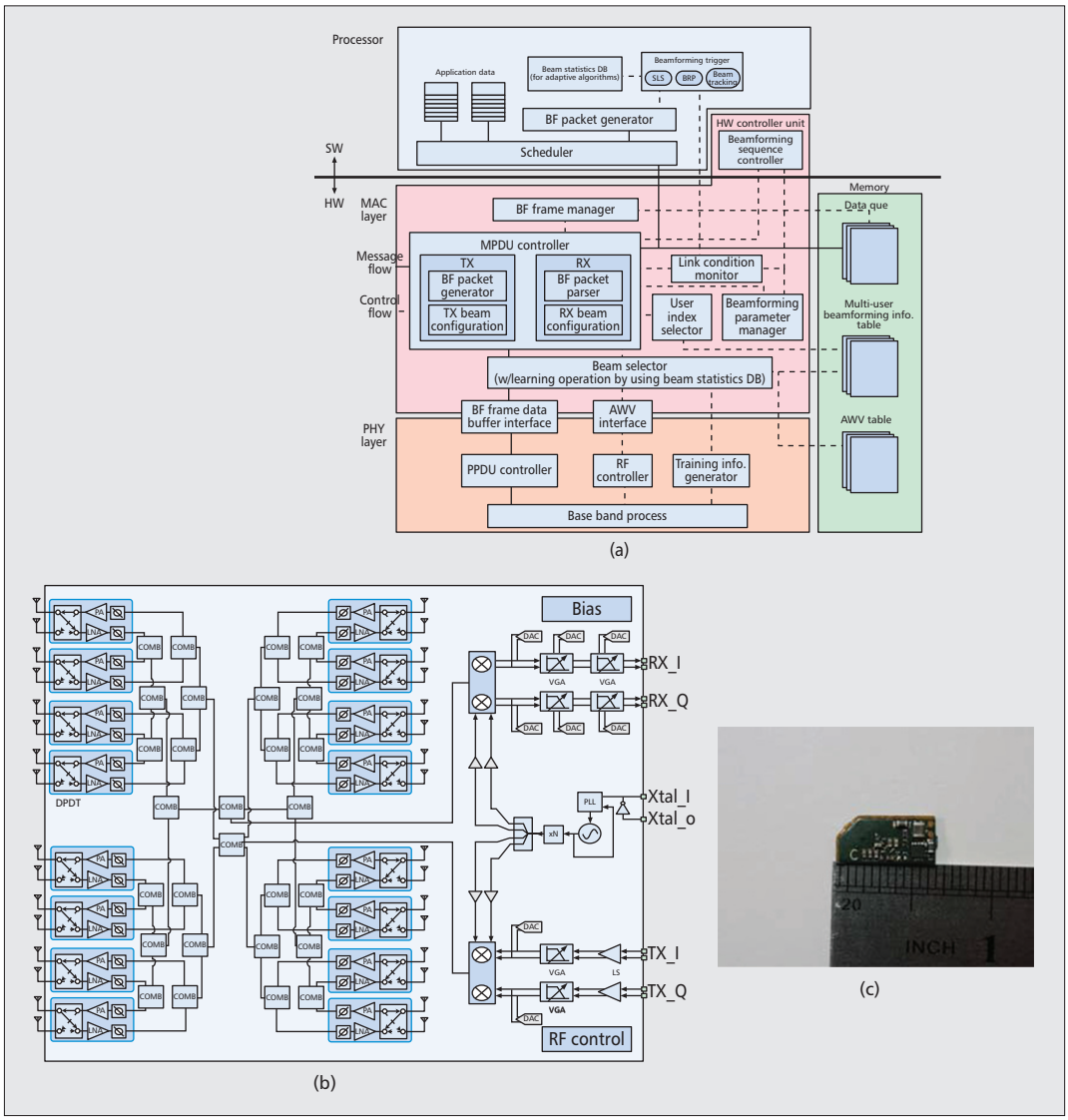


Figure 4. Advanced design and implementation of hardware-aided beamforming with respect to digital IC, RFIC, and antenna: a) novel architecture of digital IC for the implementation of fast and efficient beamforming with aids of HW; b) simplified block diagram of 60 GHz bidirectional beamforming transceiver based on DPDT switches and RF phase shifting; c) conformal phased array topology along the edges of the PCB module.

posed designs enable fast link adaptation and accurate beam training by accelerating the beam training process with the aid of hardware. Also, it provides a framework in which the enhanced algorithms can be efficiently facilitated.

Digital

As we all know, there is always a trade-off between software (SW) implementation and hardware (HW) implementation: flexibility in implementation vs. promptness in processing. This rule is applied to the implementation of beamforming as it is. In this subsection, we present a novel system architecture optimizing performance while focusing on beamforming. This includes the well collaborating functionalities of HW and SW from the perspective of the digital IC.

As stated earlier, fast and accurate link recovery is the most significant attribution of guaranteeing QoS. Simultaneously, it is one of the most challenging problems in the implementation of beamforming. In the proposed design, essential functionalities are included in the shape of digital circuits for boosting the beamforming process and obtaining the advantages of HW implementation. Handling the beamforming process in HW is very challenging. However, it returns great benefits when the hurdles are overcome.

Figure 4a is the block diagram that HW contributes efficiently to supporting and managing beamforming processes. As shown in Fig. 4a, a generous portion of core functions for beamforming is located in HW. In this subsection, we simply explain representative benefits of the HW-aided beamforming architecture for providing insight.

Immediate Link Detection — In the proposed architecture, degradation of link quality can be promptly detected using the link condition moni-

tor module in HW. This module is in charge of directly alerting the processor that beam training is necessary based on the information obtained from PHY and upper layers. As a result, the processor can immediately begin one of the beam training processes to recover the wireless link. Consequentially, the adaptation time is minimized from the moment the link condition is degraded until the beamforming process is triggered.

Fast Beam Training Process — Once the beamforming process is carried out, the training results should be reported to the peer station and applied to the system configuration as soon as possible. In the proposed design, the station can respond to the peer station promptly upon receiving a beamforming request frame. A MAC protocol data unit (MPDU) controller module implemented in HW includes unique digital circuits that can generate and parse the beamforming frames autonomously without intervention of the processor. In addition, the training results can be reflected promptly in the HW by using TX/RX beam configuration in the MPDU controller module upon receiving feedback frame from the peer station or training information from the PHY layer. This effectively results in reduction of the process time required for beamforming protocols. For example, in the proposed design, the inter-frame space for request and response during BRP iteration can achieved as low as 3 μs, which is the minimum inter-frame space time, even though the standard allows it to go up to 40 μs. As a result, the total training time consumed for BRP iteration can be reduced by more than 40 percent.

Controllability and Flexibility — The controllable parameters(beam coverage, beam width, number of beams, antenna/RF mode, etc.) used during beamforming processes are managed in the HW for dynamic configuration with the aid of the beamforming parameter manager module. Also, the information in beam statistics database is accessible for the beam selection module to facilitate QoS enhancement algorithms. In addition, we devise the HW controller unit, which is a light sub-processor designed only for micro-control of HW. Therefore, efficient beamforming operations such as sequence control for dynamic BRP iteration as well as parameter control for QoS enhanced algorithms are configurable with SW while guaranteeing flexibility.

Multi-Link Beam Switching — The proposed design enables fast beam switching for multi-link beamforming. The history of beam training for each link is managed in the memory area in the form of a table. Also, the training results are matched to the MAC address and association ID (AID) of the peer user. When the user index selector module identifies the peer user in the table, the antenna weight vectors (AWVs) for transmission and reception are updated immediately to the appropriate values corresponding to the identified user without any configuration or additional training procedures.

RF

From the perspective of RF, the wide bandwidth at unlicensed 60 GHz band available worldwide is suitable for short-range multi-gigabit-per-second high data rate wireless communications. However, to take advantage of these high data rates, the non-line-of-sight link due to the shadowing or narrow antenna beamwidth associated with the increased antenna gain should be overcome, contrary to legacy 802.11a/b/g/n, which employ omnidirectional antennas [10]. In addition, wireless communication range and beamforming coverage can be limited by polarization due to different antenna direction and the quasi-omni radiation characteristics of an mmWave antenna. In this article, a multi-chain RF beamforming architecture with a wide range of coverage is presented in order to overcome the non-line-of-sight link and extend coverage. A simplified block diagram of the fully integrated 60 GHz beamforming transceiver is shown in Fig. 4b. To achieve lower-cost 60 GHz beamforming, a direct conversion beamforming architecture based on RF path phase shifting is used. Compared to local oscillator (LO)-path and (intermediate frequency IF)/baseband-path phase shifting architectures, RF-path phase shifting results in easy extension to multiple-chain and lower-power consumption [8]. Furthermore, a double pole double through (DPDT) switch is employed in each transmitting/receiving chain to support two antennas with different coverage or polarizations. The DPDT switch is connected to two separate antennas and dual RF front-end ports, one port for a transmitter and the other for a receiver. Small size passive type 2/4-way power combiners/dividers are used for bidirectional operation with DPDT switches as shown in Fig. 4b. These switches and RF phase shifters are simultaneously controlled to generate a highly sophisticated antenna array beam that can forward in certain directions. The number of RF chains in transmitting and receiving modes can be configured from 16 to 1 to support various wireless environments. The I/Q mismatch calibration based on pre-distortion and post-distortion is employed to achieve high I/Q mismatch accuracy to levels below –35 dBc. In addition, a novel fast DC offset calibration based on digital assisted multiple current digital-to-analog converter (DAC) is employed to support IEEE 802.11ad single carrier modulation. The 60 GHz LO generation block consisting of a voltage controlled oscillator/phase locked loop and a frequency multiplier, and LO quadrature signal generator are used for both double-balanced quadrature mixers. A quadrature signal generator based on a low pass filter and high pass filter is designed for broadband and low phase/amplitude mismatch.

Antenna

For 60 GHz-based future Wi-Fi applications, antenna design methodology requires a massive makeover based on two notable criteria:

- Enabling a high antenna gain while retaining spherical beam steering coverage
- Maximizing the number of antennas within very confined real estate

The proposed design enables fast beam switching for multi-link beamforming. The history of beam training for each link is managed in the memory area in the form of a table. Also, the training results are matched to the MAC address and association ID of the peer user.

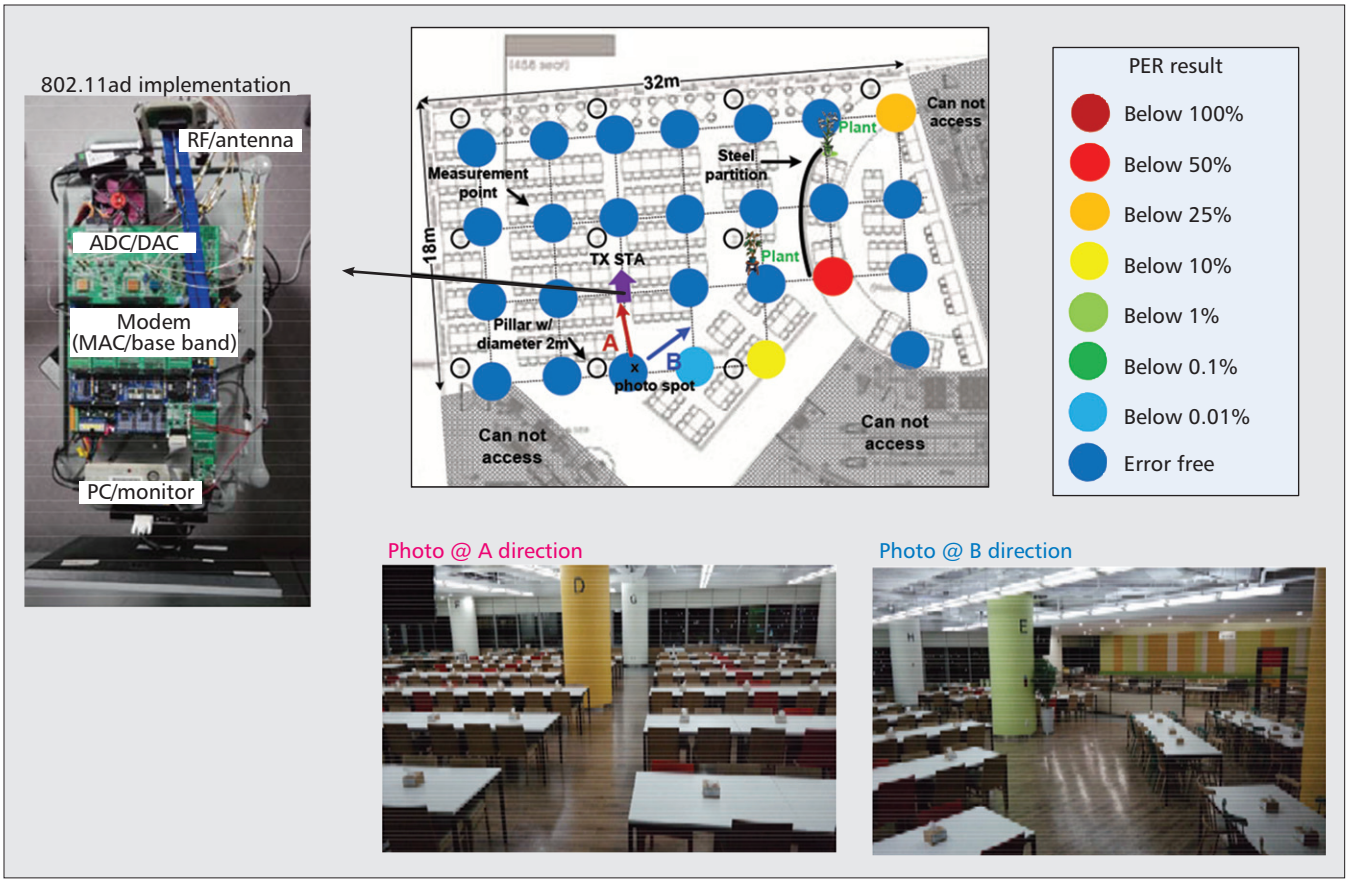


Figure 5. Environments and results of beamforming test with the IEEE 802.11ad implementation.

From the perspective of an antenna/RF engineer the aforementioned criteria are far from intuitive and can potentially be viewed to be nearly paradoxical. Due to the relatively high signal attenuation incurred by substrates and chassis at 60 GHz in comparison to that at legacy bands, future Wi-Fi modules are projected to be placed near the outer or exposed regions of the wireless device. Hence, the antenna and RFIC locations will likely be implemented in close proximity to the edge regions of the PCB (printed circuit board). We capitalize on this unique location dependency of the antenna and RFIC and confront the first challenge associated with trade-off between the high gain and the wide beam steering coverage by devising a planar antenna array that fully utilizes all faces of the PCB. As illustrated in Fig. 4c, multiple antenna elements are arranged in a conformal phased array topology along the edges of the PCB module. The slanted angle further increases the physical limitations of the beam steering range by more than 30 percent in comparison to conventional planar type phased array designs. The measured gain is confirmed to be more than 16 dBi at boresight. The second challenge related to maximizing the number of antennas while limiting the overall real estate is mitigated by designing a novel patch antenna element that is vertically oriented as opposed to the conventional horizontally oriented patch antenna. As a result, the designed footprint is reduced by more than 50 percent in comparison to previously reported designs [11, 12] with similar design conditions.

System-Level Performance

In this section, we have devised an experiment to show that beamforming technology is going to be enabling the future Wi-Fi. Using 802.11ad testbeds that we implemented with the enhanced algorithms and the advanced design specifications, indoor field test was carried out in a cafeteria located at Samsung Electronics headquarter office R3 Building 3rd floor in Suwon, South Korea.

System Parameters

The testbed consists of a modem, ADC/DAC, RF/antenna module, and PC/monitor as shown in Fig. 5. The detailed key parameters are described in Table 1. In this experiment, the available bandwidth is limited to 0.52 GHz due to the field programmable gate arrays (FPGAs) that are not affordable at very high clock frequencies. We used Xilinx Vertex-6 for the implementation of the testbed, and the back-end process for application-specific integrated circuit (ASIC) implementation is ongoing. The carrier frequency is 60.48 GHz, which corresponds to Channel 2 in the IEEE 802.11ad standard. For PER measurement, we used quadrature phase shift keying (QPSK) for modulation, low density parity check (LDPC) 3/4 for channel coding, and 2000 bytes for packet length. We also configured a reasonable effective isotropic radiated power (EIRP) value of 26 dBm.

Beamforming Test Results

In a beamforming-based system such as IEEE 802.11ad, the coverage of a single beam is very limited by the directionality. This is because of the energy concentration for overcoming the high attenuation of the high carrier frequency signals. Therefore, to assert the beamforming-based IEEE 802.11ad system as a strong candidate for future Wi-Fi, we need to confirm that it can be a superior solution while covering the same or a wider area compared to the traditional Wi-Fi for indoor applications. The test environments are shown in Fig. 5. The location and direction of the TX station (data source) are fixed, while the RX station (data sink) is roaming through the cafeteria to measure packet error rate (PER). The measurement points are spaced by 4.5 m horizontally and vertically. The heights of the TX and RX stations are equally configured as 1.1 m. In each measurement position, the best beam is selected from the beamforming protocols and algorithms stated in the previous chapters.

Remarkably, our 802.11ad-based testbed covers the entire indoor hall. At every measurement point located along the direction of the TX station, data packets are received 100 percent as expected. Moreover, robust communication paths are also found by using beamforming technology in all areas out of the direction of the TX station, even on the opposite side. Using this configuration, we were able to confirm guaranteed QoS experience between the TX and RX stations based on flawless streaming of uncompressed audio and video signals. Through the lessons from these results, we are convinced that the beamforming-based 802.11ad system properly operates in indoor environments, and the limited coverage angle for overcoming the challenging propagation characteristic in the high frequency can be settled by efficient beamforming technology.

Conclusions and Discussions

The wireless community is in the midst of an ongoing major evolution. Wide-scale enhancements ranging from data rate, reliability, and compatibility are necessary to continuously sustain the ever increasing growth of the consumer electronics industry. The IEEE 802.11ad standard has been established and is set for market readiness. However, we are still faced with numerous practical limitations that can potentially cause detrimental effects on the effectiveness of beamforming technology at the system level during mass market 60 GHz user scenarios. As presented in detail in this article for the first time, an active adaptable beamforming algorithm is vital to accurately accommodate the various surrounding environments for future Wi-Fi applications. All major layers constituting the 60 GHz hardware are specifically designed to fully enable the aforementioned beamforming technique. The effectiveness of the hardware design is corroborated by empirical results. The compelling measurement observations further ascertain the applicability and potential of IEEE 802.11ad and beamforming technology.

Key parameters	Values
Carrier frequency	60.48 GHz (Channel 2)
Bandwidth	0.52 GHz (FPGA)
Modulation	QPSK
Coding	LDPC 3/4
Packet length	2000 bytes
Effective isotropic radiated power	26 dBm

Table 1. System parameters.

Future work remains in bridging IEEE 802.11ad with legacy Wi-Fi [13, 14]. Complementary coexistence will lead to the maturity of future Wi-Fi. And the evolution of Wi-Fi will continue creating new and innovative applications.

References

[1] C. Cordeiro, D. Akhmetov, and M. Park, "IEEE 802.11ad: Introduction and Performance Evaluation of the First Multi-Gbps Wi-fi Technology," *Proc. 2010 ACM Int'l. Wksp. Mmwave Commun.: From Circuits to Networks*, Sept. 2010, pp. 3–8.

[2] E. Perahia and M. X. Gong, "Gigabit Wireless LANs: An Overview of IEEE 802.11ac and 802.11ad," *ACM SIGMOBILE Mobile Comp. and Commun. Rev.*, vol. 13, no. 3, July 2011, pp. 23–33.

[3] H. Singh *et al.*, "A 60 GHz Wireless Network for Enabling Uncompresseed Video Communication," *IEEE Commun. Mag.*, vol. 46, no. 12, Dec. 2008, pp. 71–78.

[4] H. T. Friis, "A Note on a Simple Transmission Formula," *Proc. IRE 34*, May 1946, pp. 254–56.

[5] 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.

[6] S. K. Yong, P. Xia, and A. Valdes-Garcia, *60 GHz Technology for Gb/s WLAN and WPAN: From Theory to Practice*, Wiley, 2011, pp. 17–57.

[7] S. H. Wu *et al.*, "Robust Hybrid Beamforming with Phased Antenna Arrays for Downlink SDMA in Indoor 60 GHz Channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, Aug. 2013, pp. 4542–57.

[8] A. Hajimiri *et al.*, "Phased Array Systems in Silicon," *IEEE Commun. Mag.*, vol. 42, no. 8, Aug. 2004, pp. 122–30.

[9] IEEE P802.11ad, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer(PHY) Specifications Amendment 3: Enhancement for Very High Throughput in the 60 GHz Band," Sept. 2012.

[10] S. Wyne *et al.*, "Beamforming Effects on Measured mm-Wave Channel Characteristics," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, Nov. 2011, pp. 3353–59.

[11] A. Lamminen, J. Saily, and A. R. Vimpari, "60-GHz Patch Antennas and Arrays on LTCC with Embedded-Cavity Substrates," *IEEE Trans. Antennas Propagat.*, vol. 56, no. 9, Sept. 2008, pp. 2865–74.

[12] Y. P. Zhang *et al.*, "Integration of Slot Antenna in LTCC Package for 60 GHz Radios," *Elect. Lett.*, vol. 44, no. 5, Feb. 2008, pp. 330–31.

[13] H. Singh et al., "Green Operation of Multi-Band Wireless LAN in 60 GHz and 2.4/5 GHz," *Proc. 2011 IEEE Consumer Commun. and Networking Conf.*, Jan. 2011, pp. 787–92.

[14] C. Cordeiro and S. B. Trainin, "Method, Apparatus and System for Fast Session Transfer for Multiple Frequency Band Wireless Communication," U.S. patent 13/432066, Dec. 2013.

Biographies

Ohyun Jo (ohyun.jo@samsung.com) received his B.S., M.S., and Ph.D. degrees in electrical engineering from the Korean Advanced Institute of Science and Technology (KAIST) in

2005, 2007, and 2011, respectively. Since 2011, he has been with Samsung Electronics DMC R&D Center in charge of research and development for next generation wireless communication systems. He has authored and co-authored 17+ papers, and holds 50+ patents. His research interests include wireless communication networks, protocols, and services.

WONBIN HONG (wonbin.hong@samsung.com) is currently a principal engineer at Samsung Electronics in Korea, responsible for the research and development of mmWave antennas and RF circuit design for next-generation wireless systems. He has authored and co-authored more than 40 peer reviewed journal and conference papers in the field of microwave engineering and wireless communication. He received his Ph.D. degree from the University of Michigan, Ann Arbor in 2009.

SUNG TAE CHOI (sungt.choi@samsung.com) received his B.S. degree in electronics from Kyungpook National University in 1995, and M.S. and Ph.D. degrees in mechatronics from Gwangju Institute of Science and Technology, Korea, in 1997 and 2004, respectively. In 2004, he joined the millisys, Inc., Korea. From 2005 to 2007, he was an expert researcher at the National Institute of Information and Communications Technology (NICT), Japan. Since May 2007, he has been with Samsung Electronics DMC R&D Center. His research interests include microwave and mmWave CMOS circuits/antenna design/beamforming systems/fiber-radio access systems, short-range radar systems, and multi-Gigabit wireless communication systems.

SANGHYUN CHANG (s29.chang@samsung.com) received his B.S. degree in electrical engineering from Seoul National University in 2000. He earned his M.S. and Ph.D. degrees in electrical engineering at the University of Southern California (USC) in 2002 and 2007, respectively. He was with the California Institute of Technology as a postdoctoral scholar from 2007 to 2011, jointly working at USC as a research associate from 2009 to 2011. Since 2011, he has been with Samsung Electronics DMC R&D Center. As a principal engineer, he works primary on next generation Wi-Fi standardization and development. His research interests include wireless communication, ranging, and radar system design.

CHANGYEUL KWEON (cy.kwon@samsung.com) received his B.S. degree in electronic engineering from Hanyang University, Seoul, Korea, in 1994 and his M.S. degree in electrical engineering from Stevens Institute of Technology, New Jersey, in 1996, respectively. In 2002, he joined Samsung Electronics, and is currently developing wireless solutions based on WLAN technology as a principal engineer. His current research interests are in the area of WLAN/WPAN with an emphasis on QoS guarantee.

JISUNG OH (jisung0714.oh@samsung.com) received B.S. and M.S. degrees in control and instrumentation engineering from Seoul National University in 1994 and 1996, respectively. He obtained his Ph.D. degree in electrical engineering from Stanford University, California, in 2005. Since 1996 he has been with Samsung Electronics, Suwon, Korea, and is head of the Communications Solution Lab at DMC R&D Center. He was actively involved in terrestrial and cable digital TV system development from 1996 to 2001. Currently, he participates in research and development on next generation wired/wireless communication. His research interests include multiuser/multi-antenna communications, next generation WLAN/WPAN systems, and communications ASIC design.

KYUNGWHOON CHEUN (kw.cheun@samsung.com) received his B.S. in electronics engineering from Seoul National University in 1985. He earned his M.S. and Ph.D. degrees from the University of Michigan, Ann Arbor in 1987 and 1989, respectively. He was an assistant professor at the University of Delaware from 1989 to 1991, and joined the Pohang University of Science and Technology in 1991, where he is currently a full professor. Aside from his academic duties, he served as CTO for Pulsus Technologies Inc. during 2004 to 2011, a Qualcomm partner company. Since 2012, he has been with Samsung Electronics DMC R&D Center as a senior vice president and leads the Communications Research Team in the area of next generation cellular and Wi-Fi networks.

Power Efficiency: The Next Challenge for Multi-Gigabit-per-Second Wi-Fi

Sridhar Rajagopal

ABSTRACT

Wi-Fi systems have been continuously evolving to provide a 10-fold increase in data rates every five years, exploiting higher frequencies and larger bandwidths. With each evolution, there has been an improvement in power efficiency along with data rates. This power efficiency improvement is becoming harder to sustain with increasing data rates, and power efficiency is suggested in this article as the next big challenge for multi-gigabit-per-second Wi-Fi systems. The impact of increased bandwidth on the baseband, mixed-signal, and RF design of Wi-Fi systems is studied. A low-power architecture suitable for large-bandwidth Wi-Fi systems is proposed to address the power efficiency challenge. This architecture scales with bandwidth and adapts bandwidth to the data rate requirements.

INTRODUCTION

What is next for Wi-Fi? Wi-Fi, which started as a wireless local area network connectivity technology in 1999, has become a mainstream technology invading all consumer electronic markets, with over several billion chipsets sold worldwide. Improvements in Wi-Fi have focused on multiple aspects in terms of throughput, QoS, topology, security, and so on, and Wi-Fi systems will continue to evolve in these dimensions. This work focuses on providing power efficiency relative to throughput for the next evolution of Wi-Fi. Wi-Fi started with IEEE Standard 802.11b providing 11 Mb/s in 1999, evolving to the current IEEE 802.11ac/IEEE 802.11ad standards, providing close to 7 Gb/s in 2012, a factor of 500 times improvement in data rates over the last 13 years and with a 10-fold increase in data rates every five years. This data rate increase has been enabled by increased spectrum allocation and techniques such as orthogonal frequency-division multiplexing (OFDM, IEEE 802.11a), multiple-input multiple-output (MIMO, IEEE 802.11n), multiuser MIMO (MU-MIMO, IEEE 802.11ac) and millimeter-wave non-line of sight (mmWave NLOS) beamforming at 60 GHz (IEEE 802.11ad). With new applications such as gigabit-per-second mobile data offloading, cloud computing, and the explosion in mobile video traffic with high-/ultra-high-definition (HD/UHD) content, we can expect this data rate trend of 10-fold increase every five years to continue in the future [1]. This trend is also supported by advances in hardware such as flash memory storage for mobile devices [2]. Figure 1 shows the Wi-Fi evolution timeline between 1999 and 2013 in terms of data rate, bandwidth, and carrier frequency. If this trend continues, by 2020 we can expect data rates in the range of 10–100 Gb/s from Wi-Fi networks [2].

However, the capacity or energy (in Watt hours per kilogram) of batteries used in mobile devices has improved by only 5–10 percent every year, a factor of 4 times in 15 years. This implies that we cannot depend on battery technology improvements to reduce power consumption. An order of magnitude increase in data rates will require an order of magnitude improvement in energy efficiency. Figure 2 shows the power efficiency (in bits per nanojoule) of various generations of Wi-Fi chipsets vs. data rates (in gigabits per second). A power budget of 1 W is assumed based on existing chipset implementations [3–6]. The power consumption plotted in the figure can vary between vendor implementations and should be viewed simply as a representative example. The data for single-antenna (1 × 1) implementations such as IEEE 802.11a/g, IEEE 802.11n (1 × 1), and IEEE 802.11ac (1 × 1) was obtained from [4]. The data for IEEE 802.11ad implementation was obtained from [5]. While MIMO is more spectrally efficient and is power efficient from a capacity standpoint, its impact on chip power consumption is not clear. However, a 25 percent energy efficiency improvement has been reported for 2 × 2 MIMO for an IEEE 802.11ac implementation that provides a two-fold peak data rate improvement [6]. In order for a Wi-Fi technology to be feasible for mobile implementations, it has to lie to the left of the 1 W budget line (< 1 W region) in Fig. 2. The following observations can be made from Fig. 2. First, the power efficiency has been improving with every generation with increased data rates. Second, any increase in data rates with a future generation requires an equivalent increase in power efficiency. However, it is hard to sustain

The author is with Samsung Research America — Dallas.

0163-6804/14/$25.00 © 2014 IEEE

this power efficiency trend as we continue to increase data rates for the following reasons:

• The physical layer design for achieving link capacity for a given bandwidth has been well researched, and implementations are getting close to achieving capacity [7]. It is becoming increasingly difficult to provide increased data rates at lower frequencies in the 2.4 and 5 GHz bands with limited available bandwidth of a few hundred megahertz. IEEE 802.11ac, for example, provides support for low density parity check (LDPC) channel coding, 256-quadrature amplitude modulation (QAM), eight MIMO spatial streams, downlink MU-MIMO, and 160 MHz of spectrum [8]. Further improvements in channel coding, modulation, MIMO, and increasing bandwidth at these lower frequencies to provide the next leap in data rates, while maintaining the same power efficiency, will be challenging.

• A significant contributor to the data rate increase has been the use of increased bandwidth by channel bonding from 20 to 160 MHz at lower frequencies. There is a growing focus on exploring even larger bandwidths and mmWave frequencies such as 60 GHz to provide the next leap in data rates [8]. At these higher frequencies, there are challenges related to mmWave RF design efficiency in order to be power efficient [9]. The power efficiency challenge at large bandwidths and mmWave frequencies is already visible in Fig. 2, where IEEE 802.11ad implementations at 60 GHz provide the highest power efficiency but barely meet the 1 W budget for feasible implementations.

Hence, the next leap in data rates will require us to focus on power efficiency in order to meet the requirements of mobile implementations, exploring power-efficient solutions for larger bandwidths and mmWave spectrum.

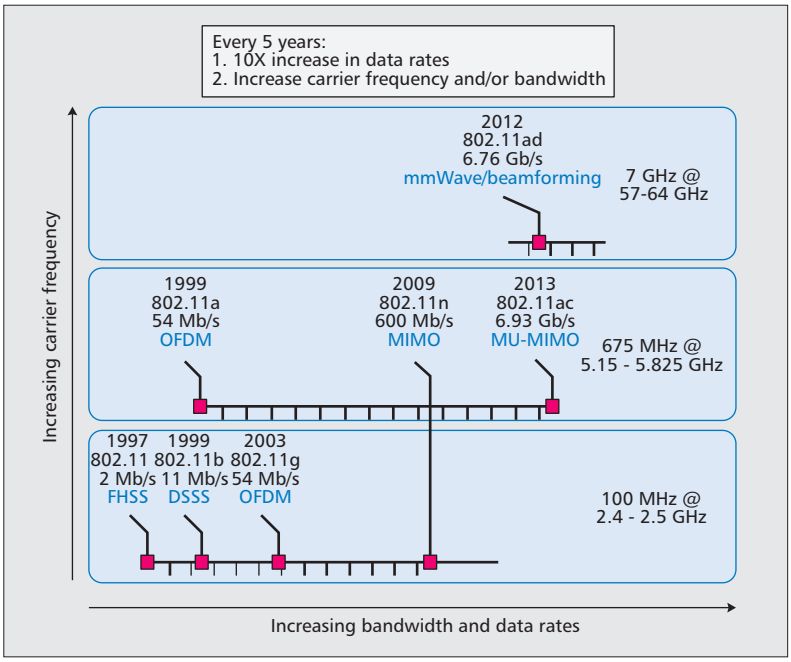


Figure 1. IEEE 802.11 (Wi-Fi) evolution timeline. The years mentioned are specification release dates from http://www.ieee802.org/11/Reports/802.11_Timelines.htm. Trend extrapolated ignoring 802.11a. Future systems will continue to provide increased data rates with more bandwidth at higher carrier frequencies.

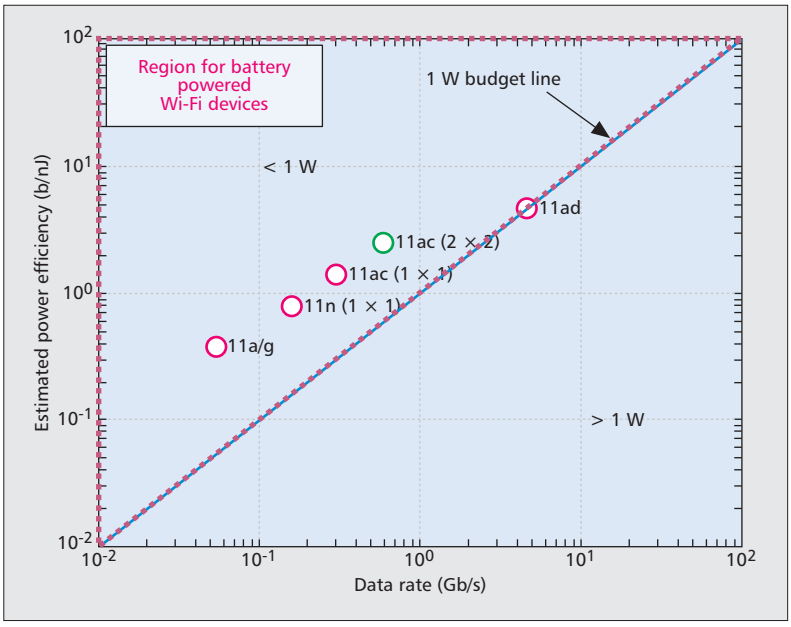


Figure 2. Power efficiency (bits per nanojoule) vs. data rates for various generations of Wi-Fi chipsets. Power efficiency increases with data rates, but this trend is becoming harder to sustain.

Impact of Increased Bandwidth on Power Efficiency

This section discusses the impact of increased bandwidth on the power efficiency of the digital/baseband, analog/mixed-signal, and RF portions of a Wi-Fi chipset.

Digital/Baseband Design

From a digital logic perspective, increased bandwidth implies faster clocks and/or more logic in order to process the increased data into the system. For digital complementary metal oxide semiconductor (CMOS) circuits, the power consumption is given by $P = \alpha CV^2 f$, where α is the switching activity, C is the capacitance, V is the voltage, and f is the operating clock frequency. Most digital circuits can also run at lower voltages when the clock frequency is reduced. If we assume a linear relationship between frequency and voltage, making $P = \alpha Cf^3$, power becomes a cubic function of operating frequency. Thus, keeping the clock frequency low is key for low power consumption in digital circuits. Hence, to meet the data rate requirements, design of parallel algorithms is critical. The assumption made in this analysis is that there are no physical area constraints in a digital design since Moore's law enables us to continue putting more transistors on a digital CMOS chip with technology node scaling. As shown in Fig. 3a, in terms of digital baseband parallelism, there can be two types:

- *Natural:* This parallelism occurs when blocks in the design are replicated multiple times. For example, a MIMO system is implemented typically using two fast Fourier transfom (FFT) blocks in parallel; this can be considered an example of natural parallelism.
- *Unexposed:* This parallelism exists within a

block but requires some work in order to be exploited. For example, to exploit parallelism within an FFT block, the data needs to be re-ordered between the intermediate stages within an FFT.

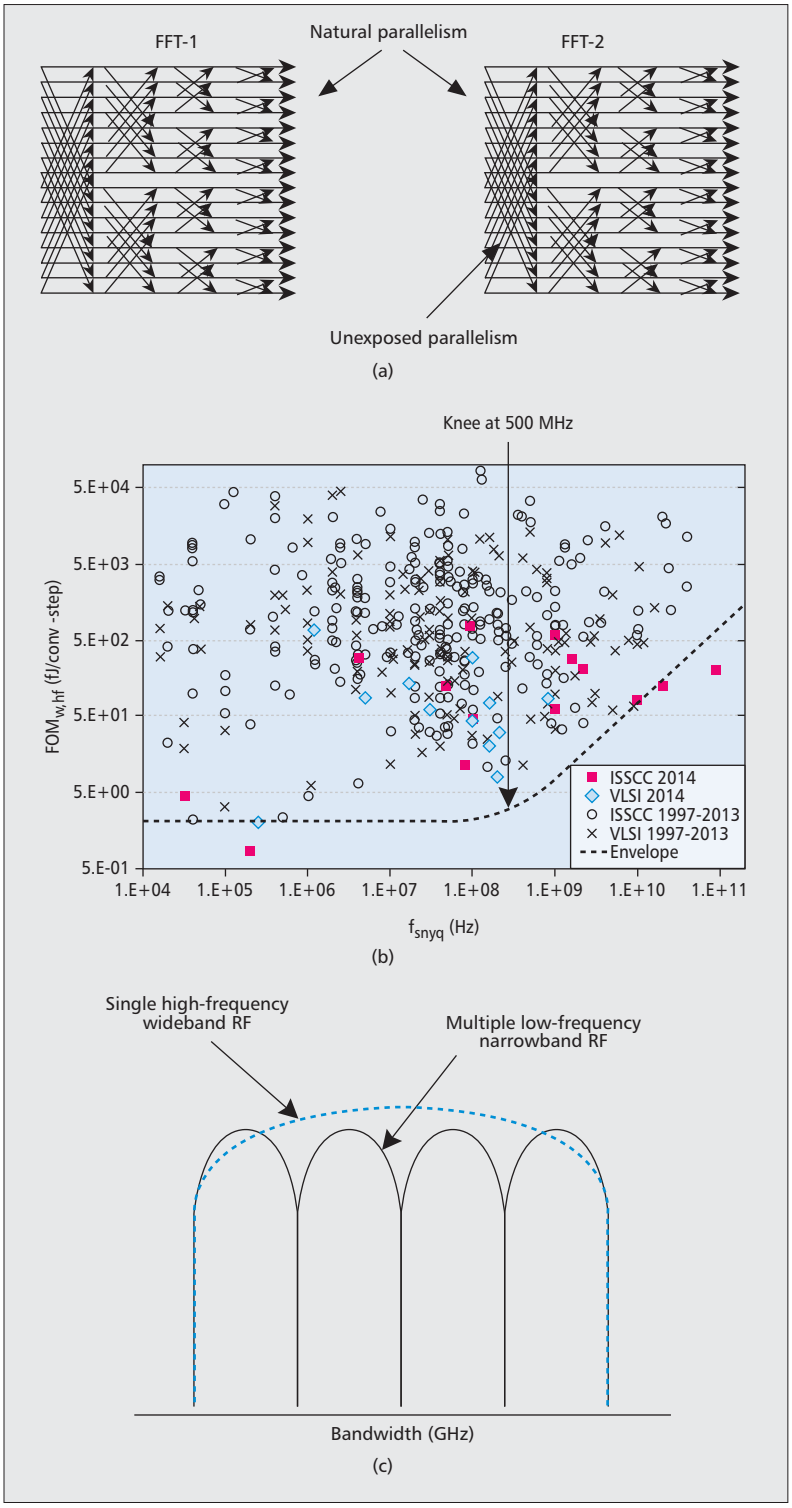


Figure 3. Impact of increased bandwidth on digital and RF circuits. Digital circuits require significant parallelism for low power, analog circuits require parallelism above a certain sampling frequency (500 MHz today), and RF circuits prefer single wideband channels: a) digital/baseband — as parallel as possible for low power; b) analog/mixed-signal — parallel above 500 MHz for low power [12]; c) RF/antenna — single wideband RF for low power.

Some blocks are easier to parallelize than others and are preferred from a power standpoint, given equivalent performance. For example, an LDPC decoder is more easily parallelized than a Turbo decoder and may be preferred in terms of a low-power implementation, assuming equivalent data rate and error rate performance, leading to its growing adoption in upcoming multi-gigabit-per-second Wi-Fi standards such as IEEE 802.11ac and 802.11ad [8].

Analog/Mixed-Signal Design

Analog and mixed signal circuits scale poorly with process technology in terms of size and speed compared to digital circuits. Large bandwidths for analog/mixed-signal designs imply that high-speed analog-to-digital converters (ADCs) and digital-to-analog converters (DACs) need to be designed that are power efficient. The power consumption of high-speed ADCs is dependent on the bit precision, sampling frequency, and other factors such as spurious free dynamic range. For high-resolution converters, the majority of the power is dissipated by circuits that drive capacitors with sizes set directly by thermal noise constraints. At larger bandwidths, ADCs become less efficient and can contribute to a significant portion of the total chip power consumption [10]. The current trend in ADCs is that power dissipation has halved every 2.5 years. This is in part due to technology scaling; however, an additional factor is the increasing use of digital error correction and calibration. Previously, non-idealities such as gain errors, offset, and distortion that were deterministic were compensated in the analog domain. However, the trend has increasingly shifted to digital correction techniques, providing power saving advantages [11]. In order to compare ADCs, a Walden figure of merit (FOM_W) is typically used that normalizes the power consumption P by the sampling frequency f_s and effective number of bits (ENOB), given by $FOM_W = P/f_s 2^{ENOB}$. Figure 3b shows the FOM_W for ADCs published at ISSCC and VLSI conferences between 1997 to 2014 [12]. From the envelope curve showing the lower bound on ADC power efficiency vs. sampling frequency, it can be seen that power efficiency of ADCs is constant up to a certain Nyquist sampling frequency f_{snyq} (around 500 MHz in this plot) but starts degrading above this frequency due to loss in sampling efficiency [10, 11]. Hence, in order to be power efficient, the system should be designed such that ADCs do not require sampling beyond 500 MHz bandwidth assuming today's ADC technology. The current knee at 500 MHz may improve further over time with advances in ADC technology but at a much slower rate compared to improvements in digital process technology.

RF/Antenna Design

For RF circuits, the design is dependent on the fractional bandwidth, which is the ratio of the bandwidth to the carrier frequency. The fractional bandwidth is also related to the Q-factor $Q = f_c/\Delta_f$, where Δ_f is the 3 dB bandwidth and f_c is the carrier frequency. A 10 percent ratio for the fractional bandwidth is typical. At low frequencies, antennas and RF circuits traditionally operate on smaller bandwidths, and hence can have good Q and efficiency. As we increase bandwidth for the next leap in data rates, designing a wideband RF

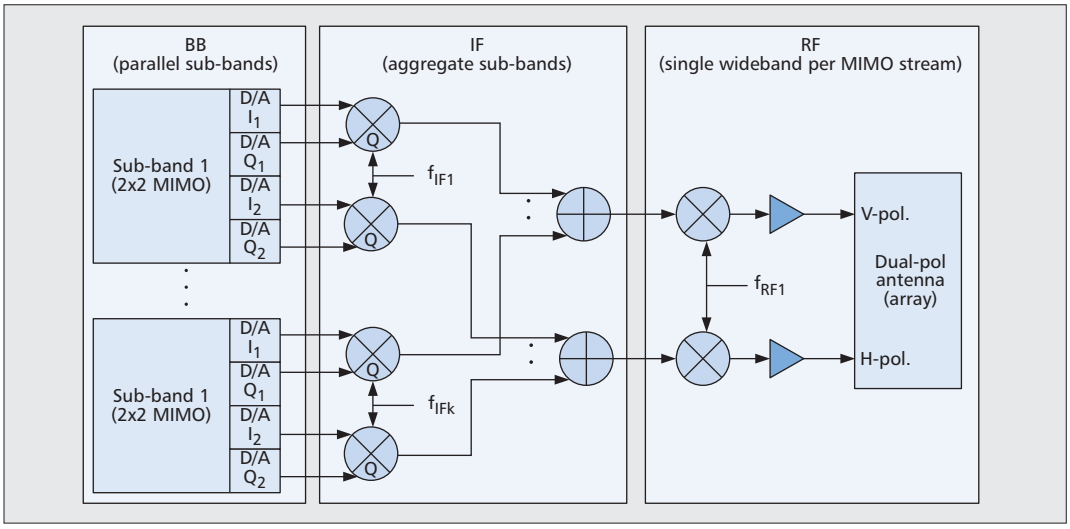


Figure 4. Low-power architecture for large-bandwidth Wi-Fi systems. Multiple sub-bands at lower bandwidth are aggregated at IF and upconverted to a single wideband RF per MIMO stream. The receiver follows a similar architecture to the transmitter shown in the figure.

The key concept here is to simplify analog and RF circuits for high data rates and large bandwidths to save power and compensate for the impairments by adding complexity on the digital baseband side, where the digital circuits can scale with technology process nodes.

and antenna front-end in order to attain good Q factor and efficiency over the entire band will be challenging at low frequencies, potentially requiring multiple sub-band RF circuits in order to maintain performance. Multiple sub-band RF circuits will come at the price of increased power consumption, size, and cost, especially when MIMO support is considered. In order to provide larger bandwidth, higher frequency operation should be considered, where a single RF can provide the increased bandwidth. Since RF circuits do not scale well with process technology in terms of area, single wideband RF at high frequencies can also help reduce area and cost compared to multiple narrowband RF circuits at lower frequencies to support the same bandwidth, as shown in Fig. 3c. However, RF circuits at higher frequencies have poor efficiency due to low gain transistors, lossy interconnects, and low maximum supply voltage [9], which should be factored in the design considerations.

Low-Power Architecture for Large-Bandwidth Systems

Figure 4 shows the proposed low-power architecture design for large-bandwidth systems that considers the impact of increased bandwidth on the baseband, mixed signal, and RF circuits. The frequency band is broken into multiple sub-bands, each not exceeding 500 MHz in bandwidth. MIMO techniques can be performed within each sub-band. The sub-bands are then aggregated at intermediate frequency (IF) and sent to a single wideband RF. This concept can be applied to large-bandwidth Wi-Fi systems, such as IEEE 802.11ad, which uses 2 GHz channelization at 60 GHz. Each 2 GHz band can be broken, for example, into 500 MHz sub-bands for improved power efficiency. Similar concepts have been presented in related work to solve ADC bottlenecks for large-bandwidth systems [10, 13]. However, this work also evaluates, in addition to the ADC, the impact of increased bandwidth on the baseband, IF, and RF parts in the architecture design. The impact of this architecture in a system during operation is also presented, demonstrating power savings in idle listening and active modes of communication.

Digital/Baseband Design

The total bandwidth is broken into multiple sub-bands, each supporting a maximum of 500 MHz of bandwidth. Standard modulation techniques such as MIMO, OFDM, and single-carrier modes can be applied within each sub-band. Each sub-band has the flexibility for independent operation if required, so the hardware corresponding to unused sub-bands can be turned off to save power. The medium access control (MAC) can provide multiple parallel streams of data from/to the application layer for each sub-band in order to support 10–100 Gb/s data rates. The sub-bands are placed adjacent to each other with minimum guard spacing. This leads to increased interference at the band edges that can be compensated at the receiver using digital signal processing techniques for interference cancellation. Techniques for cancelling intercarrier interference (ICI) across sub-bands as well as intersymbol interference (ISI) within a sub-band due to channel dispersion were studied in [13]. Furthermore, imperfect mixers can generate harmonics in adjacent sub-bands, and these harmonics need to be compensated via digital signal processing. Calibration for ADCs can also be done on the digital side for low power operation [11]. The key concept here is to simplify analog and RF circuits for high data rates and large bandwidths to save power and compensate for the impairments by adding complexity on the digital baseband side, where the digital circuits can scale with technology process nodes.

Analog/Mixed-Signal Design

High-speed ADC design approaches [11, 14] consist of three steps:

- *Signal decomposition:* where the input analog signal is sub-sampled in time or frequency

As we explore multi-band Wi-Fi solutions in 2.4, 5, and 60 GHz (and possibly even higher frequencies), smart system design techniques need to be applied in order to share hardware resources and architect the system in the most power-efficient manner.

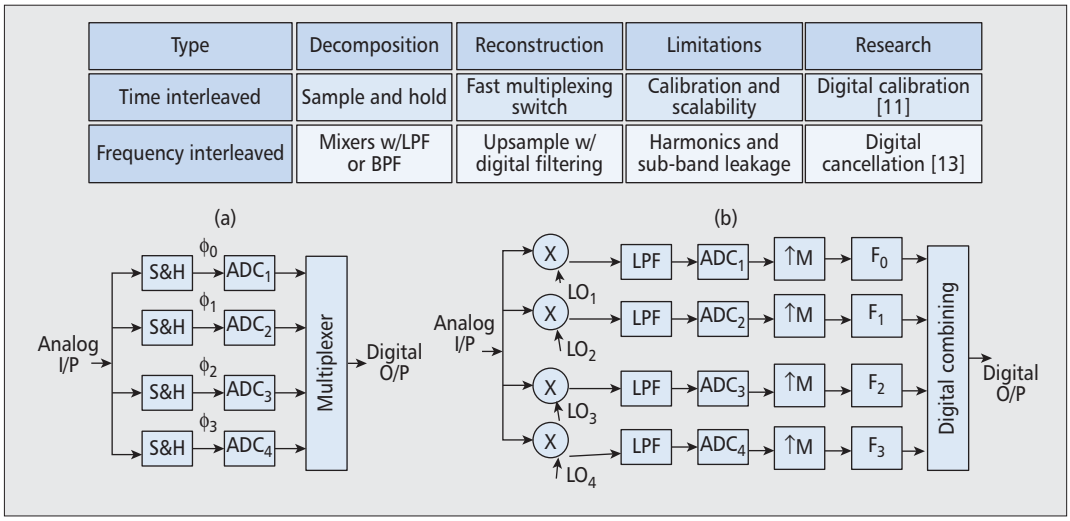


Figure 5. Flexible bandwidth support for a high-speed ADC using time-interleaved or frequency-interleaved sub-ADCs. Sub-ADCs can be turned "OFF" to flexibly adapt bandwidth to the application requirements and save power: a) time-interleaved ADC; b) frequency-interleaved ADC.

- *Quantization:* where each sub-sampled analog signal is quantized into a digital signal
- *Reconstruction:* where the full bandwidth digital signal is reconstructed from the sub-sampled signal

As we explore large-bandwidth systems, typically ADCs using multiple low-power sub-ADCs are employed in order to scale the operational bandwidth [11, 14]. These sub-ADCs can be interleaved in time or frequency. For example, a time-interleaved ADC [11] samples and holds the signal at different phases of the clock, quantizes each signal, and then uses a fast multiplexing switch to reconstruct the full bandwidth signal. The limitations of time-interleaved ADC are the need to calibrate offset errors, gain mismatches, and timing skews. This calibration is done digitally in order to benefit from Moore's law. Frequency-interleaved ADCs operate by splitting the signal into narrower sub-bands using mixers and low pass filters or using band pass filters and undersampling [14]. In this case, there can be leakage between adjacent sub-bands, and the harmonics of one band can fall into adjacent bands that require digital cancellation [13]. The trade-offs of time- and frequency-interleaved ADCs are summarized in Fig. 5. Note that the use of multiple sub-ADCs for such high-speed ADC architectures can be used to provide flexible bandwidth and save power by turning off sub-ADCs in time or frequency when required.

IF/RF/Antenna Design

A new IF stage is required to aggregate the multiple sub-bands that can be upconverted to a single wideband RF per MIMO stream. This IF stage adds overhead compared to a traditional full-band mixer. However, the complexity and power consumption of mixers (usually implemented using a nonlinear device, e.g., a few transistors) is minimal compared to the cost of large-bandwidth ADCs. If image suppression due to harmonics is not sufficient in baseband, there may be a case to add harmonic reject mixers. Even in this case, the complexity increase has been considered minimal [15]. The RF and antennas are designed to cover the entire bandwidth of operation. As shown in Fig. 4, a dual-polarized antenna structure is considered such that two-stream MIMO can be operated even under LOS conditions for Wi-Fi using polarization as an additional degree of freedom for MIMO operation. At mmWave frequencies such as 60 GHz, an antenna array could be considered to provide sufficient gain using analog/RF beamforming.

Flexible Bandwidth Modes of Operation

A natural way to adapt the ADC bandwidth during communication is to adapt the input clock (down-clocking). This provides a linear reduction of power consumption with frequency. However, this requires additional settling time for the phase locked loop (PLL) and synthesizers, during which synchronization could be lost. For time and frequency interleaved systems, a more efficient way could be to turn off sub-ADCs as needed to adapt the bandwidth to the requirements of the application. We assume there are n sub-ADCs, and the flexible bandwidth modes use integer multiples of f_s/n bandwidth, where f_s is the original Nyquist sampling frequency of the ADC. The flexible bandwidth modes use k out of n sub-ADCs, keeping the remaining $(n - k)$ sub-ADCs "OFF." In this case, the PLL/clock for the sub-ADCs does not need to be adapted. Note that "OFF" does not necessarily mean "no voltage" but rather a "standby or sleep" state where sub-ADCs can be turned "ON" quickly without recalibration. This can be achieved, for example, by clock gating the sub-ADCs. The interleaving overhead is implementation-dependent, but we can assume a 10 percent overhead based on [8] for the sub-ADCs in the OFF mode.

Figure 6 shows ADC operation during idle listen and active states in the flexible bandwidth mode. In this flexible bandwidth mode, only one

sub-ADC is kept ON during idle listening, where the Wi-Fi device is constantly looking for unpredictably arriving packets or assessing a clear channel. This can have significant power savings in the idle listening state, where high-speed ADCs may be the dominant source of power consumption since only 1 of *n* sub-ADCs is ON while the rest are OFF. Having received initial acquisition on one sub-band, more sub-ADCs can turned ON for data channel operation depending on the application requirements. The synchronization performance of such a system has been studied in [16], showing 33 percent power reduction during idle listening. The packet detection, AGC adaptation, and beamforming training can be done within a single sub-band. Any control information for authentication and beaconing can also make use of this sub-band. The data channel can then be adapted to use multiple sub-bands depending on the requirements of the application, based on the configuration set in the control channel transmission. This bandwidth flexibility also allows support for increased channelization, which can be useful for dense Wi-Fi networks.

Conclusion

This article serves as a tutorial article to highlight the growing need to address power consumption in future multi-gigabit-per-second Wi-Fi systems. The article shows that the requirement of improving power efficiency at the same rate as throughput is becoming increasingly harder to obtain as we continue scaling data rates for future systems. The article explores the impact of increased bandwidth on power consumption in the digital, mixed signal and RF parts of the physical layer. A low-power architecture design is proposed for large-bandwidth Wi-Fi systems that can scale with bandwidth and adapt to the data rate requirements of the application.

There are still several challenges that need to be solved for power-efficient designs for future Wi-Fi systems. As we explore larger bandwidths at mmWave frequencies and beyond, techniques to improve power efficiency of RF circuits at higher frequencies must be investigated. Another important consideration is the need for backward compatibility and coexistence. Newer system designs for power efficiency should consider backward-compatible and coexistence techniques in order to coexist with existing designs sharing the same unlicensed spectrum. As we explore multi-band Wi-Fi solutions in 2.4, 5, and 60 GHz (and possibly even higher frequencies), smart system design techniques need to be applied in order to share hardware resources and architect the system in the most power-efficient manner.

Acknowledgments

The author is grateful to his colleagues at Samsung Research America — Dallas for their valuable feedback and discussions.

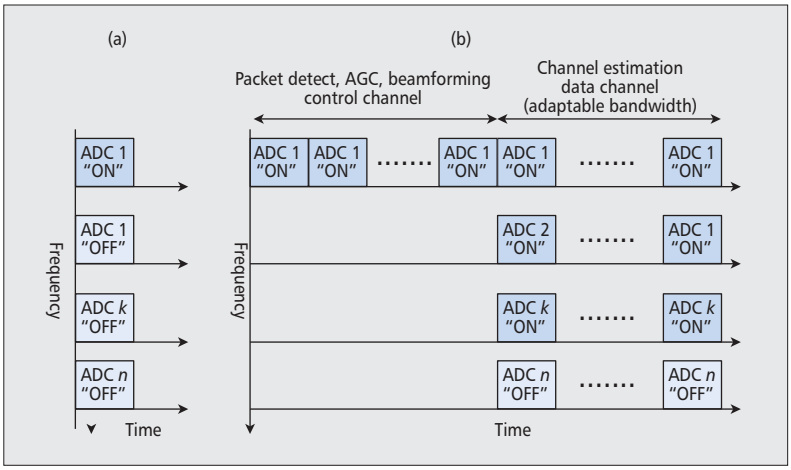


Figure 6. System operation in idle and active states with flexible bandwidth adaptation. Only one of the sub-bands is kept active until data transmission occurs to save power. a) idle listening; b) active.

References

[1] L. Garber, "Wi-Fi Races into a Faster Future," *IEEE Computer*, vol.45, no.3, Mar. 2012, pp. 13–16.

[2] G. Fettweis, F. Guderian, and S. Krone, "Entering the Path towards Terabit/s Wireless Links," *Proc. Design, Automation & Test in Europe Conf. & Exhibition*, Mar. 2011, pp. 1–6.

[3] G. P. Perrucci, F.H.P Fitzek, and J. Widmer, "Survey on Energy Consumption Entities on the Smartphone Platform," *Proc. IEEE VTC-Spring*, May 2011, pp. 1–6.

[4] Unex Technology Corp.; http://unex.com.tw/802-11ac-gigabit-wi-fi.

[5] M. Boers *et al.*, "A 16TX/16RX 60GHz 802.11ad Chipset with Single Coaxial Interface and Polarization Diversity," *Proc. IEEE ISSCC*, Feb. 2014, pp. 344–45.

[6] Broadcom, "Broadcom Introduces First 5G Wi-Fi 2x2 MIMO Combo Chip for Smartphones," press release at Mobile World Congress, Feb. 2014.

[7] M. Dohler *et al.*, "Is the PHY Layer Dead?" *IEEE Commun. Mag.*, vol. 49, no. 4, Apr. 2011, pp. 159–65.

[8] L. Verma, M. Fakharzadeh, and S. Choi, "Wi-Fi on Steroids: 802.11ac and 802.11ad," *IEEE Wireless Commun.*, vol. 20, no. 6, Dec. 2013, pp. 30–35.

[9] A. M. Niknejad and H. Hashemi, Eds., *mm-Wave Silicon Technology: 60 GHz and Beyond*, Springer, 2008.

[10] J. Singh, S. Ponnuru, and U. Madhow, "Multi-Gigabit Communication: The ADC Bottleneck," *Proc. IEEE Int'l Conf. Ultra-Wideband*, Sep. 2009, pp. 22–27.

[11] D. Stepanovic, "Calibration Techniques for Time-Interleaved SAR A/D Converters," UC Berkeley tech. rep. no. UCB/EECS-2012-225, Dec. 2012.

[12] B. Murmann, "ADC Performance Survey 1997–2014," http://www.stanford.edu/~murmann/adcsurvey.html.

[13] H. Zhang, S. Venkateswaran, and U. Madhow, "Analog Multitone with Interference Suppression: Relieving the ADC bottleneck for Wideband 60 GHz Systems," *Proc. IEEE GLOBECOM*, Dec. 2012, pp. 2305–10.

[14] G. Ding *et al.*, "Frequency-Interleaving Technique for High-Speed A/D Conversion," *Proc. IEEE Int'l. Symp. Circuits and Sys.*, vol.1, May 2003, pp. 857–60.

[15] V. Dyadyuk *et al.*, "A Multigigabit Millimeter-Wave Communication System with Improved Spectral Efficiency," *IEEE Trans. Microwave Theory and Tech.*, vol. 55, no. 12, Dec. 2007, pp. 2813–21.

[16] S. Rajagopal, E. Pisek, and S. A-Surra,"Low Power Synchronization Design for Large Bandwidth Wireless LAN Systems," accepted for publication, *Proc. IEEE GLOBECOM*, Dec. 2014.

Biography

Sridhar Rajagopal [SM] (sridhar.r@samsung.com) is currently a senior staff engineer at Samsung Research America in Dallas, Texas. He received his M.S. and Ph.D. degrees in electrical and computer engineering from Rice University. He has previously worked at Nokia Research Center and at WiQuest Communications, and has contributed to multiple communication standards. His research interests are in algorithms and architectures for short-range, high-throughput, and low-power technologies, mmWave, and optical wireless communication.

An Advanced Wi-Fi Data Service Platform Coupled with a Cellular Network for Future Wireless Access

Riichi Kudo, Yasushi Takatori, B. A. Hirantha Sithira Abeysekera, Yasuhiko Inoue, Atsushi Murase, Akira Yamada, Hiroto Yasuda, and Yukihiko Okumura

ABSTRACT

Wireless LAN devices are now everywhere because of the rapid spread of smart wireless devices. Demand for far richer content services is also driving the expansion of mobile traffic. Converging the cellular network with Wi-Fi is a reasonable way to support the increasing mobile traffic because most mobile user terminals already have Wi-Fi interfaces. More opportunities for Wi-Fi use require further enhancement of system capacity and manageability, especially in the high-density Wi-Fi network. This is because the chronic depletion of system resources is becoming a significant problem in the Wi-Fi network given the increases in Wi-Fi density and traffic. This article introduces a Wi-Fi data service platform coupled with cellular networks, which strengthens the synergy of two networks. Enhanced monitoring and performance prediction are essential to provide a high-grade user experience in high-density Wi-Fi environments.

INTRODUCTION

The proliferation of smart wireless devices such as smart phones and tablets has been rapidly increasing mobile traffic. Demand for far richer content services is depleting radio resources. These trends underlie the explosion in mobile traffic that has been observed in actual networks. It is predicted that the mobile traffic will explode 1000-fold in the next decade if the current trend of a two-fold traffic increase each year continues. Thus, it is crucial to increase system capacity and prepare for future wireless access systems in the upcoming tremendous mobile traffic era. Converging the cellular network with Wi-Fi is a reasonable way to support increasing mobile traffic because most mobile user terminals already have Wi-Fi interfaces [1, 2]. The latest Wi-Fi, 802.11ac, already offers a maximum physical layer data rate (PHY rate) of over 400 Mb/s in the 5 GHz band even for single antenna arrangements [3]. This makes Wi-Fi attractive as one of the user data planes in cellular systems. In the Third Generation Partnership Project (3GPP), the access network domain selection function (ANDSF) [4] defines how to switch between Wi-Fi and cellular access by exchanging policies, and such interworking with cellular systems is also accelerating the use of Wi-Fi.

Figure 1 shows the relationship between the three key performance factors of Wi-Fi in future wireless access systems. The first factor is operator/service provider performance; this includes the number of Wi-Fi services supported and the amount of Wi-Fi traffic. The second one is the proliferation of Wi-Fi systems. This factor is significant for vendors. The last one is user experience. The first flow indicates that the increase in Wi-Fi traffic will require more Wi-Fi deployments. In the second flow, the increase in Wi-Fi system density improves user experience. The improved user experience/satisfaction triggers the third flow, increases in the Wi-Fi traffic. More opportunities to access the Wi-Fi network will also fuel the creation of novel network services. The technological challenge is to support the second flow, that is, how to improve user experience with high-density Wi-Fi. This issue has been intensively discussed in the IEEE 802.11-HEW Study Group and has led to the formation of the IEEE 802.11 Task Group ax (TGax) that defines both physical (PHY) and medium access control (MAC) layers to achieve improvement in the average throughput per station in dense deployment scenarios [5]. Further enhancement can be expected by introducing Wi-Fi management over the entire network; the goal is to manage not only high-density Wi-Fi systems, but also traffic load balancing among Wi-Fi systems and cellular systems.

Toward this vision, it is essential to predict Wi-Fi performance in the complex radio environment where multiple Wi-Fi access points (APs) and various wireless systems contend for the same frequency resources. The extremely dense Wi-Fi deployment and huge mobile traffic will make the Wi-Fi throughput more unstable. In this article, we advocate the deployment of a converged network architecture wherein the Wi-Fi management architecture integrates Wi-Fi network management with the cellular network links. We first review Wi-Fi performance and

Riichi Kudo, Yasushi Takatori, B. A. Hirantha Sithira Abeysekera, Yasuhiko Inoue, and Atsushi Murase are with NTT Corporation.

Akira Yamada, Hiroto Yasuda, and Yukihiko Okumura are with NTT DOCOMO INC.

0163-6804/14/$25.00 © 2014 IEEE

$$\begin{cases} R = \dfrac{E[\text{successfully received payload information in total time } T]}{E[\text{time used for data symbols successfully received}]} \\ P = \dfrac{E[\text{time used for data symbols successfully received}]}{E[\text{total time } T]} \end{cases} \quad (1)$$

describe what causes the throughput degradation in dense Wi-Fi deployment. The, we discuss the enhanced Wi-Fi radio environment monitoring and Wi-Fi performance prediction. In the following, the management of the Wi-Fi network and mobile traffic are discussed. Finally, we summarize how the positive circulation in Fig. 1 is accelerated by the described architecture.

Wi-Fi Performance in Future Wireless Access Systems

Future Wi-Fi is expected to be densely deployed and to overlay the cellular network. Effective utilization of Wi-Fi networks becomes more important to improve total mobile network capacity as well as user experience. Thus, performance estimation of wireless networks is necessary to manage network capacity and user experience. The current public/enterprise network uses a Wi-Fi access controller (AC); it improves Wi-Fi performance by managing radio resources and traffic loads among multiple APs. However, even the current AC faces difficulties in estimating Wi-Fi throughput in high-density Wi-Fi deployments because of the complex radio environment; each Wi-Fi works autonomously in unlicensed frequency bands. Wi-Fi throughput is expressed as the expectation of the payload information transmitted over total time T [6]. Total time T is the duration from when the payload information is enqueued to the receipt of an acknowledgment. Thus, Wi-Fi user throughput can be divided into PHY rate R and successful transmission ratio P as in Eq. 1, where $E[\cdot]$ denotes expectation. $R \times P$ denotes the Wi-Fi user throughput. The related factors are shown in Table 1. Estimating Wi-Fi throughput requires an assessment of all these factors. The PHY rate can be predicted more easily than successful transmission ratio P since the relating factors can be directly measured in the basic service set (BSS), which consists of a single AP and its associated UEs, and works as a basic block in Wi-Fi systems. The instability of P is mainly caused by coexistence with other BSSs and other wireless systems. The effects of the interference from such wireless systems must also be predicted to control Wi-Fi throughput.

OBSS

The influence of the overlapping BSS (OBSS) corresponds to the factors of resource sharing with other BSSs in the same channel, overlapping among multiple BSSs (exposed terminal problem) for transmission opportunity in the BSS, and overlapping among multiple Wi-Fi nodes (hidden terminal problem) for packet collision probability in Table 1. The factors are strongly related to the traffic of other BSSs, which can dynamically change. The hidden and exposed terminal problems are inherent to carrier sense multiple access with collision avoidance (CSMA/CA) protocols, and these problems are enhanced when the traffic of other BSSs occupies almost all radio resources. In the sparse traffic condition, the hidden terminal problem can be mitigated by using request to send/clear to send (RTS/CTS) handshake. This, however, becomes an imperfect solution in a heavy traffic environment since the RTS receiver cannot transmit CTS packets as it detects OBSS transmission. Furthermore, the current IEEE 802.11 wireless LAN standard does not have any solutions to the exposed terminal problem for unregulated dense AP deployment. The impact of the hidden and exposed terminal problems depends on the Wi-Fi link condition and their traffic. These problems degrade the transmission opportunity and/or increase the packet collision probability with the OBSS.

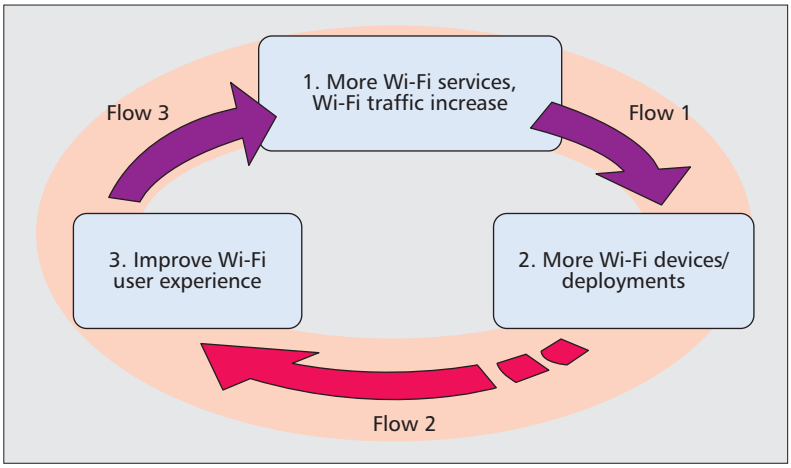


Figure 1. Positive cycle of Wi-Fi deployment.

Inter-System Interference

The inter-system interference also influences the transmission opportunity and packet collision probability (Table 1). It is well known that several interference sources exist in the 2.4 GHz band (microwave ovens, Bluetooth, etc.). For instance, microwave ovens can degrade Wi-Fi performance but only at specified times and places. The 5 GHz band offers much greater allocatable bandwidth than the 2.4 GHz band. However, large parts of the 5 GHz band are shared with radar systems, and Wi-Fi systems must avoid interfering with them. Upon discovering an active radar, the AP must terminate the current data transmission, select a new channel, inform the impacted UEs, and move to the new channel. In the dense Wi-Fi and heavy traffic environment, such a drastic change in Wi-Fi traffic may cause large-scale disconnection. Furthermore, new systems using 5 GHz unlicensed band channels are being discussed in [7]. If such new

systems become popular, the characteristic of inter-system interference must be analyzed to predict the influence of them.

Numerical Example

Figure 2 shows the cumulative distribution functions (CDF) of the Wi-Fi throughput for 200 UEs in a residential simulation scenario [8].The offered downlink traffic load for each UE was set to 5 Mb/s, 10 Mb/s, 20 Mb/s, 40 Mb/s, and 65 Mb/s to each user station (STA). Since each AP has two UEs, the load-PHY ratios, which are the ratios of load carried to the PHY rate, were 7.5 percent (= 5 [Mb/s] × 2 [UEs]/130 [Mb/s]), 15, 30, 60, and 100 percent. The Wi-Fi channel was randomly selected from among four frequency channels Ch. 36, Ch. 40, Ch. 44, and Ch. 48 included in the 5.15–5.25 GHz band; this corresponds to the condition wherein a 5 GHz radar system prevents any channel sharing. When the disadvantaged UEs are defined as those with throughput less than 5 Mb/s, it is found that the rate of the disadvantaged UEs increases as the load-PHY ratio increases. Even when the load-PHY ratio is 15 percent, UE throughput can fall to under 1 Mb/s. This shows that the Wi-Fi performance is degraded by allocation of excessive amounts of traffic. Since the PHY rate is fixed, the significant throughput instability is ascribed to ratio *P*. Figure 3 shows the links between the APs and UEs. The circles and white ellipses denote the APs and UEs. The lines between Wi-Fi nodes indicate that the nodes can hear each other's signals; that is, the signals exceed the clear channel assessment (CCA) threshold in the communication channel. It is found that large clusters with more than 20 APs are formed. The maximum number and average number of the detectable APs at each Wi-Fi node in this simulation are 13 and 7.2, respectively. In the actual Wi-Fi environment, other factors such as real-time rate adaptation and uplink traffic should also be taken into account. The wireless resource is more deeply depleted as the PHY rate decreases and uplink transmission increases, which worsens the Wi-Fi conditions.

R: PHY rate	Frequency bandwidth	Channel usage of the other systems Channel usage condition of other BSSs
	Modulation and coding scheme (MCS)	Received signal strength identification (RSSI) of the Wi-Fi link
		Spatial multiplexing condition in single-user/multiuser-MIMO
		Interference condition under carrier sense level
		Rate adaptation algorithm
P: successful transmission ratio	Transmission opportunity	DATA length of other nodes
		Resource sharing with other UEs in the same BSS
		Resource sharing with other UEs in the other BSS with the same channel (OBSS interference)
		Interference condition among multiple BSSs, that is, expose terminal problem (OBSS interference)
		Interference from other systems (inter-system interference)
		Channel changes (due to radar detection) (inter-system interference)
	Packet collision	Random access contention
		Interference condition among multiple Wi-Fi nodes, that is, hidden terminal problem (OBSS interference)
		Interference from other systems (inter-system interference)
	Packet error due to MCS mismatch	Bit error rate of the successfully transmitted packet
		Bit amount of a packet
	Payload information Efficiency	DATA length
		Control/management signal overhead
	Available backhaul capacity for Wi-Fi	Backhaul capacity
		Traffic condition of other communication systems sharing the same backhaul network

Table 1. Factors associated with Wi-Fi throughput.

Management Architecture

Instantaneous throughput degradation must be avoided for UEs that want constant throughput even in high-density Wi-Fi environments. To get Wi-Fi performance under control, we advocate the deployment of an enhanced Wi-Fi management architecture (Fig. 4). The key components are the management block, Wi-Fi network, destination UEs, data source, and cellular network. The red and black lines denote the control signal paths and user data paths. The Wi-Fi network and Wi-Fi traffic should be managed to maximize Wi-Fi system capacity and avoid Wi-Fi access failures. Wi-Fi network and traffic management are conducted by three functions in the management block. The monitoring database collects feedback from the APs and UEs via control signal paths. The APs' and UEs' feedback are obtained by way of backhaul lines and the control plane of the cellular systems, respectively. The Wi-Fi performance estimator uses the information in the monitoring database. The radio resource manager maintains the Wi-Fi network and controls the traffic using the control signal paths. The entities of the Wi-Fi network are APs, UE-MRs (mobile routers, MRs, or UEs using a tethering function). It is expected that some of the APs are controlled by the management block and have the function of monitoring the Wi-Fi radio environment. A UE-MR also acts as UE and AP because it supports the access of both cellular systems and Wi-Fi to the destination STA that has only Wi-Fi interfaces. The cellular network has various base transceiver stations (BTSs) for macrocells, small cells,

and spot cells. The BTS can collect the feedback from the destination UEs and UE- MRs, and transfer the Wi-Fi setting signals to the UE-MRs. The last block is destination UEs that include the STAs with only Wi-Fi interfaces, and UEs that have both Wi-Fi and cellular interfaces. The management block monitors the Wi-Fi radio environment, estimates Wi-Fi performance, and conducts radio resource management. Then the effectiveness of the radio resource management is evaluated by monitoring. Cycling through these three management blocks keeps improving the user experience even in the face of drastic radio environment change.

Radio Environment Monitoring

Information on the radio environment will become more critical for the success of the future Wi-Fi platform. Thus, the environment monitoring function should be enhanced in order to collect Wi-Fi information from not only APs but also UEs. Monitoring by the UE has the advantage that the Wi-Fi radio environment is monitored regardless of Wi-Fi connection status. This architecture enables wide-area radio resource management corresponding to the coverage area of BTSs. The APs and UEs can collect not only the information of their Wi-Fi links, but also the control signals from the surrounding APs and the interference from Wi-Fi and other systems. The monitoring information related to the Wi-Fi link could include RSSI, PHY rate, frame length, and frame loss rate. Other monitoring information could be the number of active Wi-Fi devices operating on each frequency channel, their capability information, MAC addresses of the transmitter and desired receiver, network allocation vector (NAV) information for virtual carrier sensing, frame type and frame length information of MAC frames, and a noise histogram, which is defined in IEEE 802.11 TGk Amendment "Access Network Query Protocol (ANQP)" in IEEE 802.11 TGu. The detected information must be associated with the Wi-Fi device ID, such as a MAC address, to estimate the Wi-Fi performance.

It is expected that the UEs will also measure the user experience parameters (i.e., throughput and/or latency) and notify the measured values to the monitoring database. For Wi-Fi throughput assessment, PHY rate and successful transmission ratio should be measured to comprehend the radio environment in Wi-Fi systems. The measured successful transmission ratio can be compared to the estimated one at the Wi-Fi performance estimator. When there are gaps between the measured and estimated throughputs, the Wi-Fi performance estimator must be improved to reduce the gap. The throughput measurements will be conducted periodically or after radio resource management activity. Efficient throughput measurement must be studied to alleviate the load of the control signal paths.

Wi-Fi Performance Estimation

The Wi-Fi performance estimator evaluates the successful transmission ratio in addition to the PHY rate estimation. The former should be evaluated in two aspects: OBSS in Wi-Fi systems and inter-system interference from other systems. Since the impact of OBSS is strongly related to the Wi-Fi links (Fig. 3), the Wi-Fi performance estimator extracts link information from the information in the monitoring database. The successful transmission ratio can be evaluated from estimated Wi-Fi links and assumptions of the Wi-Fi load PHY ratios of the links. We show an example of successful transmission ratio evaluation based on the residence model [8]. The Wi-Fi performance estimator uses the perfect Wi-Fi link information and feasible traffic assumptions. Figure 5 shows the evaluated successful transmission ratio of a certain BSS corresponding to channel selection. It was assumed that there is only downlink transmission, and the load-PHY ratios of the other BSSs are randomly distributed between 0 to 100 percent, and the load-PHY ratio for the focused BSS was set to 100 percent. The Wi-Fi links in four available channels (Ch. 36–48) are also shown in Fig. 5. The number of other detected APs is two for all channels. However, the distributions of the successful transmission probabilities are different in each channel. The successful transmission ratio in Ch. 36 is greater than those in other channels. The 10 percent outage of the successful transmission probabilities is 0.43 in Ch. 36. If one UE communicates with the AP in Ch. 36 with a PHY rate of 54 Mb/s, the Wi-Fi throughput is expected to be greater than 22 Mb/s with a probability of 90 percent. On the other hand, the successful transmission ratio in Ch. 48 could be less than 0.1 percent. The accuracy of the performance evaluation can be improved by using uplink/downlink traffic models of the other BSSs. Since

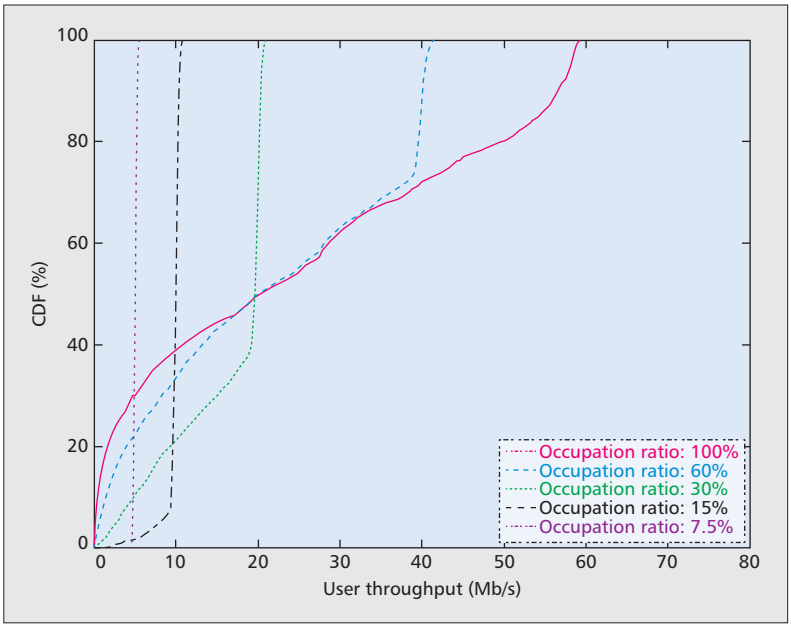


Figure 2. Numerical example of user throughputs. The PHY rate of the point-to-point transmission was set to 130 Mb/s (2 spatial streams, 64-QAM, 5/6 coding rate, 800 μs GI), transmission power of all Wi-Fi devices was set to 17 dBm, and only downlink transmission was considered. The building has 100 rooms each of which has one Wi-Fi AP and two UEs randomly located. The number of available channels is 4 by assuming the coexistence of a severe radar source.

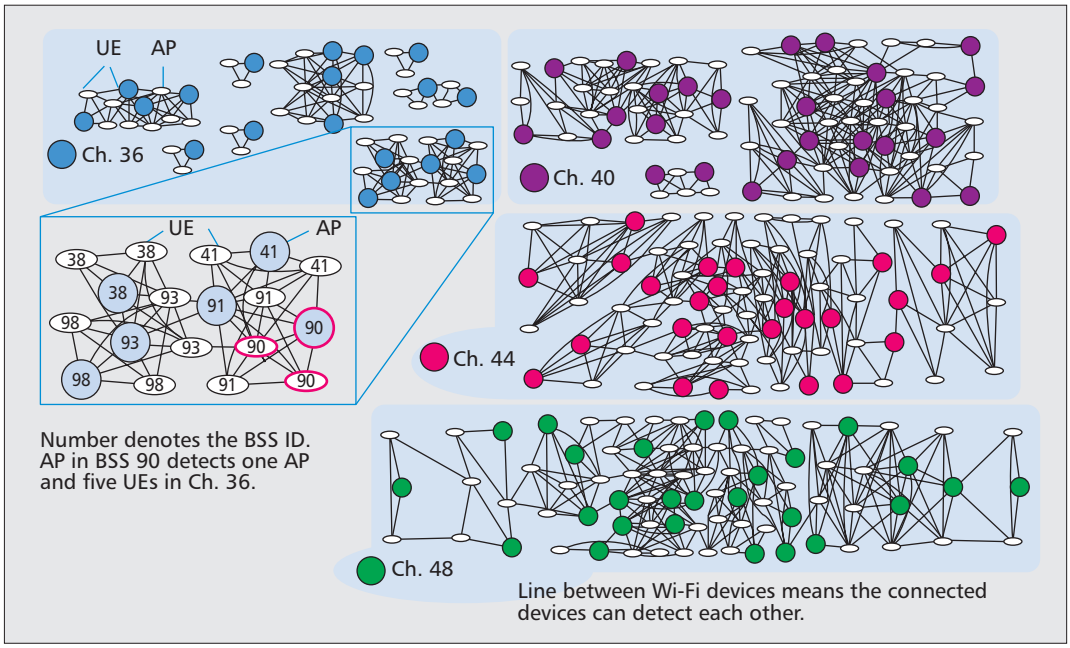


Figure 3. Wi-Fi links of a snapshot in the residence scenario simulations where the 100 APs randomly select their channels.

the traffic varies depending on the time or day, accurate traffic models improve the reliability of the successful transmission ratio. The calculation load of the successful transmission ratio evaluation can be reduced by simplifying the consideration of MAC layer operation. For example, the links that are likely to be affected by the hidden or exposed terminal problems can be identified by comparing IDs detected by Wi-Fi devices.

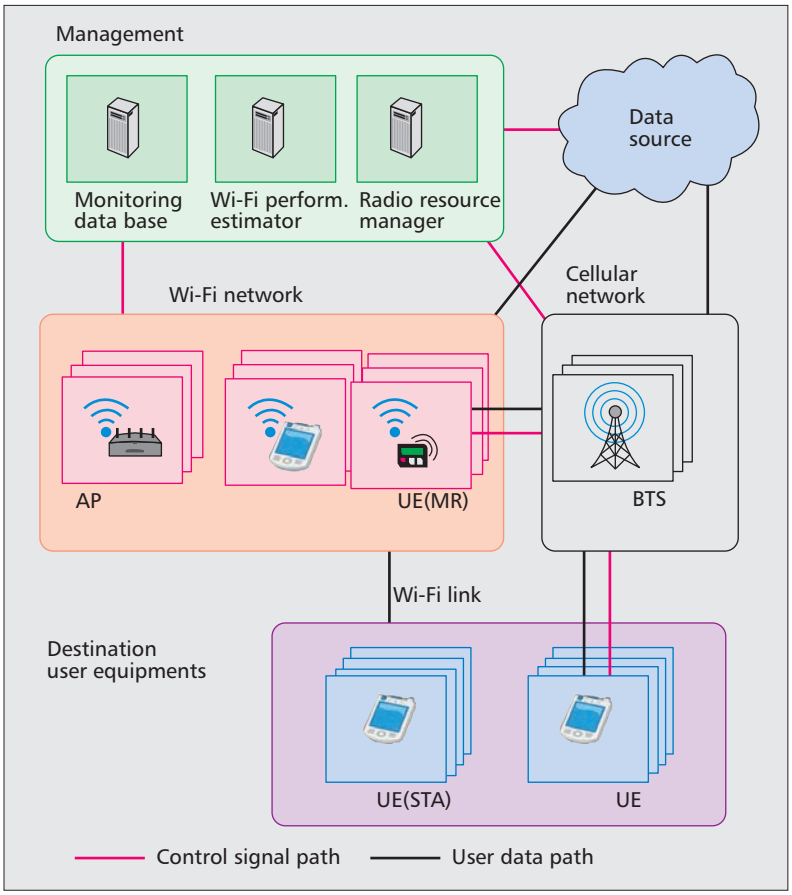


Figure 4. Wi-Fi management platform in the mobile network.

In addition to the static Wi-Fi links, the AP configuration changes, and the occasional Wi-Fi links caused by the UE-MRs also influence Wi-Fi performance. The Wi-Fi estimator needs to consider possible AP configurations. In the environment where the appearance of UE-MRs degrades the Wi-Fi performance significantly, feasible Wi-Fi links need to be considered in the successful transmission ratio evaluation. Thus, the Wi-Fi performance estimator prepares several possible Wi-Fi links to calculate Fig. 5.

The successful transmission ratio against the inter-system interference is also evaluated by using the monitored information. To evaluate the successful transmission ratio in a similar way to OBSS interference, the characteristics of the inter-system interference and detectable locations such as Wi-Fi links must be measured. For example, the incident rate of significant inter-system interference can be multiplied by the successful transmission ratio for OBSS interference. The Wi-Fi performance estimator analyzes the positions of detecting Wi-Fi devices, channels, and characteristics of the detected time. The actions of the APs that detect radar should also be analyzed in the 5 GHz band to evaluate the impact of the channel switching of the other APs. The influence of the channel switching of the other APs detecting radar can be evaluated as the successful transmission ratio against OBSS interference.

Radio Resource Management

The functions of the radio resource manager are AP configuration management and traffic management. AP configuration management deter-

mines the AP configuration parameters of APs and UE-MRs, and detects the abnormality in the Wi-Fi network to maximize and maintain Wi-Fi network capacity. To improve the successful transmission ratio, we need to analyze the transmission power, primary channel selection, channel aggregation use, contention window size, and RTS/CTS use. CCA threshold control is possible to develop to construct the adequate Wi-Fi links. Traffic management supports UE access selection or determines adequate UE access to avoid access failures caused by unstable Wi-Fi performance.

AP Configuration Management

The AP configuration is determined to improve the successful transmission ratio by considering possible parameters in the Wi-Fi performance estimator. When evaluating the successful transmission ratio as shown in Fig. 5, the AP configuration set the channel of the AP to Ch. 36. It is possible to choose the channel to maximize arbitrary percent outage or average data packet occupation ratios. Since the Wi-Fi radio environment information is enhanced by the monitoring via the control plane of the cellular network, the UE (STA) with only Wi-Fi interface also gains an advantage with radio resource management. If the monitoring database detects the radio environment reaction, a change in AP configurations of the surrounding APs after setting the AP configuration, the AP configuration management needs to consider the further AP configuration change corresponding to the changed radio environment. The timing of the AP configuration change must be also optimized by considering the load of the AP configuration change.

Since the Wi-Fi links shown in Fig. 3 impact Wi-Fi performance, it is possible for the AP configuration management to consider which links should be added or eliminated. The Wi-Fi links should be constructed to improve the successful transmission ratio. The position of an added AP can be roughly estimated from the results of radio environment monitoring at UEs and their position information. The Wi-Fi links created by UE-MRs should be managed by the AP configuration management to minimize the system capacity decrease created by the Wi-Fi link of UE-MRs. When the communication of the UE-MR decreases the Wi-Fi performance of some links, the radio resource management can reduce the cellular traffic of the UE-MR.

Traffic Control

The evaluated Wi-Fi performance can be used for traffic control to optimally share the mobile traffic among Wi-Fi and cellular network access. The wireless access of the UE with Wi-Fi interface must be determined from among Wi-Fi access, cellular access, and link aggregation with both. The access determination can be realized by three approaches: Wi-Fi performance information sharing, access policy distribution, and access designation notification. In performance information sharing, the evaluated Wi-Fi performance of the existing AP is reported to the UE. The UE can consider the user condition/preference, power saving mode, CPU and memory condition, application information, and cellular data traffic limit contract. Policy distribution yields a loosely coupled Wi-Fi/cellular network by using the performance evaluation in both Wi-Fi and cellular systems. The policy needs to be extended to support the distribution of the successful transmission ratio. In the final approach, the UE access is determined by the mobile , and the UE uses the designated access. In this approach, the UEs do not care which wireless access medium they are using. Since the Wi-Fi access failure must be avoided in order not to damage the user experience, robust Wi-Fi access and/or adaptive link aggregation are essential.

In the tremendous mobile traffic era, the radio resources of most wireless systems will approach complete depletion. Excessive offloading to a Wi-Fi network may collapse the Wi-Fi network. Smart integration of Wi-Fi and cellular systems is required. Link aggregation between Wi-Fi and cellular systems is one promising approach to use the Wi-Fi network without damaging the user experience. Link aggregation can be implemented by several schemes using common IP addresses in both Wi-Fi and cellular networks or multiple IP addresses (i.e., S2a mobility based on general packet radio service [GPRS] tunneling protocol [GTP] and WLAN access to the enhanced packet core [EPC] network via SaMOG or MPTCP) [9]. Since Wi-Fi through-

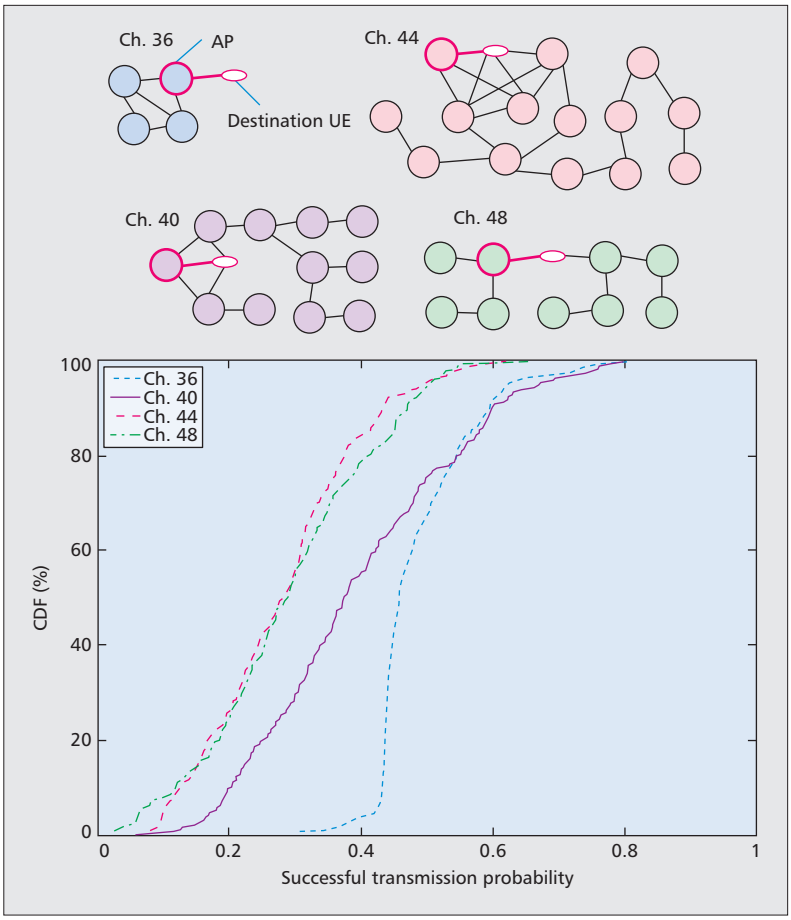


Figure 5. Cumulative distribution function of successful transmission probabilities of four available channels. It is assumed that the load-PHY ratios of the other BSSs are set to from 0 to 100 percent. Only downlink transmission using RTS/CTS exchanges is considered.

The user experience is expected to be well maintained by effectively combining the significant but erratic Wi-Fi capacity with the steady cellular throughput.

put can be unstable, it is important for mobile operators to minimize cellular traffic for link aggregation UEs based on the successful transmission ratio of the Wi-Fi link.

The probability distribution of the successful transmission ratio is useful in selecting Wi-Fi access, cellular access, or link aggregation. The probability of extremely low Wi-Fi throughput is a useful indicator of Wi-Fi throughput instability. The information on successful transmission ratio may enable traffic allocation that is suboptimal but avoids unexpected throughput degradation. The user experience is expected to be well maintained by effectively combining the significant but erratic Wi-Fi capacity with the steady cellular throughput.

Conclusion and Future Directions

It is expected that the interworking of Wi-Fi and cellular systems will more fully realize the potential of Wi-Fi systems. This article has investigated the Wi-Fi data service platform in future wireless access systems given the anticipation of huge mobile data traffic loads. It was shown that the actual throughput of Wi-Fi is unstable in dense Wi-Fi deployment scenarios with the huge mobile traffic due to the complex Wi-Fi radio environment and interference from other wireless systems in unlicensed frequency bands. A possible Wi-Fi management architecture was introduced to effectively utilize the potential capacity of Wi-Fi in the mobile network. Monitoring the radio environment at APs and UEs enables the Wi-Fi data service platform to estimate the successful transmission ratio by considering the OBSS influence and inter-system interference. It is expected to strengthen UE adoption of monitoring functions to achieve better access. The control plane in cellular systems gets rid of the Wi-Fi monitoring limitation and further improves the Wi-Fi radio environment information.

Extension of the architecture with the following three approaches has the potential to achieve further enhancement. The first approach is consideration of more wireless systems, including convergence with other wireless access in different frequency bands as well as coexistence with other wireless systems operated by different operators. The second one is coordination with backhaul/fronthaul networks. The bottleneck of backhaul/fronthaul networks is expected to be identified by total traffic monitoring, and the calculation resources can be distributed into various access networks and nodes. The third one is cross-layer optimization taking into account the suitable application layer structure for the network architecture described in this article.

The first approach, more wireless systems, requires performance evaluations of various wireless access schemes based on propagation models and use cases. Even if the throughput of a single wireless system is unstable, the combination of wireless accesses provides a high-grade user experience. To enable seamless wireless access, the values of the radio resources of the different wireless accesses must be assessed by considering the PHY rate and successful transmission ratio distribution. In addition, cooperation among multiple operators should be considered on frequency resource sharing. The game theoretic approach is a new way to analyze such scenarios from the viewpoint of the frequency resource management strategy [10].

The backhaul/fronthaul access networks can be a bottleneck as the aggregated throughput in the dense mobile UEs increases. The optimization of all traffic, including wired and wireless accesses, will be required to resolve the backhaul/fronthaul restriction. The analysis of the aggregated throughput and latency information is expected to clarify the impact of the backhaul/fronthaul capacities. High transmission capacity between a management server and UEs may realize the concept of network computing where calculation of resource offloading is achieved by distributing the calculations among various nodes (e.g., management server, AP/BTS, and mobile devices) in the total mobile network.

Cross-layer optimization is the key area to add value to future wireless access systems. Since the user can utilize multiple networks/links simultaneously, it may be good to divide an application service into multiple application modules that use different wireless accesses. Requirements (e.g., throughput and latency) for each application module are sent to the management server, and multiple rounds of network management per user are conducted to provide the suitable connection for the application service. This new approach will accelerate aggressive cross-layer optimization.

The extension with the three approaches described above requires large-scale optimization. Unfortunately, it makes it difficult to deterministically derive the optimum values for all controllable parameters. Application of machine-learning techniques may be the most practical way to obtain a quasi-optimum solution within a limited time because performance levels with various parameters for a Wi-Fi radio environment are stored at the management server and can utilize the information as big data. Furthermore, information that is not directly relevant to mobile traffic may enhance the performance estimation at the management server. For instance, information related to people's behavior at public events or in weather have the potential to improve the mobile traffic forecasts and thus enable better proactive offloading.

In the radio resource depletion condition, one-way offloading to a Wi-Fi network may damage user experience due to the throughput instability related to the inherent problems of the unlicensed band. We believe that the future Wi-Fi platform architecture for future wireless access systems will optimize traffic sharing between Wi-Fi and cellular networks, and enhance overall mobile network performance. The synergy of the two networks will further strengthen the positive circle shown in Fig. 1.

References

[1] D. Cavalcanti *et al.*, "Issues in Integrating Cellular Networks, WLANs, and MANETs: A Futuristic Heterogeneous Wireless Network," *IEEE Wireless Commun.*, vol. 12, no. 3, June 2005, pp. 30–41.

[2] W. Song, W. Zhuang, and Y. Cheng "Load Balancing for Cellular/WLAN Integrated Networks," *IEEE Network*, vol. 21, no. 1, Jan. 2007, pp. 27–33.
[3] IEEE Std for Information Technology, "Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks — Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," Dec. 2013.
[4] 3GPP, TS 24.312 (V12.2.0), "Access Network Discovery and Selection Function (ANDSF) Management Object (MO)," Sept. 2013.
[5] A. Stephens, "802.11 March 2014 WG Motions," Doc.: IEEE 802.11-14/0254r3, Mar. 2014.
[6] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE JSAC*, Mar. 2000, pp. 535–47.
[7] 3GPP RP-140057, "On the Primacy of Licensed Spectrum in Relation to the Proposal of Using LTE for a Licensed-Assisted Access to Unlicensed Spectrum," TSG-RAN #63, Mar. 2014.
[8] S. Merlin, *et al.*, "HEW SG Simulation Scenarios," IEEE 802.11-13/1001r6, Jan. 2014.
[9] J. Korhonen, *et al.*, "Toward Network Controlled IP Traffic Offloading," *IEEE Commun. Mag.*, vol. 51, no. 3, Mar. 2013, pp. 96–102.
[10] N. Nie and C. Comaniciu, "Adaptive Channel Allocation Spectrum Etiquette for Cognitive Radio Networks," *Proc. IEEE DYSPAN 2005*, Nov. 2005, pp. 269–78.

Biographies

Riichi Kudo (kudo.riichi@lab.ntt.co.jp) received his B.S. and M.S. degrees in geophysics from Tohoku University, Japan, in 2001 and 2003, respectively. He received his Ph.D. degree in informatics from Kyoto University in 2010. In 2003, he joined NTT Network Innovation Laboratories, Japan. He was a visiting fellow at the Centre for Communications Research (CCR), Bristol University, United Kingdom, from 2012 to 2013. He is now working for NTT Access Network Service Systems Laboratories. He received the Young Engineer Award from IEICE in 2006, IEEE AP-S Japan Chapter Young Engineer Award in 2010, and the Best Paper Award from IEICE in 2011.

Yasushi Takatori [M] received his B.E. degree in electrical and communication engineering and M.E. degree in system information engineering from Tohoku University in 1993 and 1995, respectively. He received his Ph.D. degree in wireless communication engineering from Aalborg University, Denmark, in 2005. He joined NTT in 1995. He is currently working on R&D of a high-efficiency wireless access platform as well as the optical core network. He has served as a co-chair of COEX Ad Hoc in IEEE 802.11ac from 2009 to 2010. He was a visiting researcher at the Center for TeleInFrastrutur (CTIF), Aalborg University from 2004 and 2005. He received the Best Paper Award from IEICE in 2011. He is a Senior Member of IEICE.

B. A. Hirantha Sithira Abeysekera [M] received B.Eng., M.Eng., and Ph.D.(Eng.) degrees in communications engineering from Osaka University, Suita, Japan, in 2005, 2007, and 2010, respectively. In 2010, he joined Nippon Telegraph and Telephone (NTT) Corporation, Japan. At present, he is working for NTT Access Network Service Systems Laboratories in Yokosuka, Japan. His research interests include design and performance evaluation of next generation wireless networks. He received the IEEE VTS Japan Student Paper award in 2009. He is a member of IEICE.

Yasuhiko Inoue [M] received his B.E. and M.E. degrees in electrical engineering from Keio University, Kanagawa, Japan, in 1992 and 1994, respectively. In 1994, he joined NTT Wireless Systems Laboratories, where he engaged in R&D of a personal handy phone (PHS) packet data communication system. Since 1997, he has been working on R&D of IEEE 802.11 wireless LAN systems and has participated in standardization activities since 2001. Currently, he is working on the research, development, and standardization of high-efficiency wireless LAN systems. He was a visiting scholar at Stanford University from 2005 to 2006. He is now working as a senior research engineer at NTT Access Network Service Systems Laboratories, Yokosuka, Japan. He received the Young Engineer Award from IEICE in 2001 and received a Contributor Award for the IEEE 802.11j standard from the IEEE Standards Association in 2004. He is currently serving as Secretary of the IEEE 802.11 TGax the High Efficiency WLAN standardization group.

Atsushi Murase received his Bachelor's degree in electronics and communications engineering from Waseda University, Tokyo, in March 1981 and joined NTT directly. He received his Ph.D. degree in cellular radio control channel design for random access control and paging signal broadcasting from Waseda University in March 1991. He has broad experience in 1G to 4G mobile communication systems development, especially base stations, controllers, and 3G FOMA terminals through more than 30 years of active work in mobile communication R&D of NTT and NTT DOCOMO. He stayed at British Telecom Labs in the United Kingdom from 1989 to 1990 as an exchange researcher. He was president and CEO of DOCOMO Communications Laboratories Europe GmbH in Munich, Germany, from 2002 to 2005. He was managing director of research laboratories, NTT DOCOMO, from 2007 to 2012. He was director of NTT Microsystem Integration Laboratories, NTT, from 2012 to 2013. He has been director of NTT Science and Core Technology Laboratory group since 2013.

Akira Yamada received his B.E. and M.S. degrees from Tokyo Institute of Technology. He joined NTT DOCOMO INC in 2000. His research topics are in the areas of 5G radio access network architecture design, traffic offloading, IEEE 802.11 standardization, public Wi-Fi access service development, and ISDB-Tmm (Japanese mobile TV system) standardization.

Hiroto Yasuda received his B.E. and M.S. degrees from the University of Electro-Communications. He joined NTT DOCOMO INC. in 2009. His research topics are in the areas of 5G radio access network architecture design, Wi-Fi standardization, traffic offloading, cellular-Wi-Fi interworking, and cognitive radio.

Yukihiko Okumura [SM] received his B.S. and M.S. degrees in electrical engineering from the Tokyo University of Science in 1989 and 1991, respectively, and his Ph.D. degree in engineering from Tohoku University in 2006. In 1991, he joined the Radio Communication Systems Laboratories of NTT, Kanagawa, Japan, and since 1992, he has been engaged in the research, standardization and development of wideband/broadband mobile radio communication technologies, terminals, and systems at the NTT Mobile Communications Network, INC. (now NTT DOCOMO INC.), Kanagawa, Japan. He is a Senior Member of the IEICE of Japan.

We believe that the advanced Wi-Fi platform architecture for future wireless access systems will optimize the traffic sharing between the Wi-Fi and cellular networks and enhance the overall mobile network performance.

Enabling the Coexistence of LTE and Wi-Fi in Unlicensed Bands

Fuad M. Abinader, Jr., Erika P. L. Almeida, Fabiano S. Chaves, André M. Cavalcante, Robson D. Vieira, Rafael C. D. Paiva, Angilberto M. Sobrinho, Sayantan Choudhury, Esa Tuomaala, Klaus Doppler, and Vicente A. Sousa, Jr.

ABSTRACT

The expansion of wireless broadband access network deployments is resulting in increased scarcity of available radio spectrum. It is very likely that in the near future, cellular technologies and wireless local area networks will need to coexist in the same unlicensed bands. However, the two most prominent technologies, LTE and Wi-Fi, were designed to work in different bands and not to coexist in a shared band. In this article, we discuss the issues that arise from the concurrent operation of LTE and Wi-Fi in the same unlicensed bands from the point of view of radio resource management. We show that Wi-Fi is severely impacted by LTE transmissions; hence, the coexistence of LTE and Wi-Fi needs to be carefully investigated. We discuss some possible coexistence mechanisms and future research directions that may lead to successful joint deployment of LTE and Wi-Fi in the same unlicensed band.

INTRODUCTION

Wireless communication infrastructure is facing a great challenge with the expanding demand for wireless broadband access to Internet. A recent forecast study [1] indicates that a traffic growth beyond 500-fold is expected between 2010 and 2020, assuming the same increase in data usage is maintained. In order to improve the capacity, the Third Generation Partnership Project (3GPP) standards group has been investigating the performance gains obtained by small cell deployment in Long Term Evolution (LTE) Release 12 and beyond. On the other hand, the IEEE 802.11 Working Group (WG) just ratified a new IEEE 802.11ax Task Group (TGax) primarily focused on enhancing the system performance of Wi-Fi in dense deployment scenarios [2].

However, some practical issues impose limitations on large-scale small cell deployments. First, there are increased costs for deploying and maintaining the required infrastructure. Customers are increasingly seeing wireless Internet access as a utility, and premium taxation on faster connections becomes less of an option since the introduction of flat rate tariffs. So, as revenue is not increasing at the same pace as expenditures [1], capacity expansion requiring larger capital expenditures (CAPEX), such as acquisition and installation of cells, and operating expenditures (OPEX) (e.g., backbone maintenance) becomes an economic challenge. The second issue relates to the diminishing availability of radio spectrum, a fundamental resource that is both finite and expensive. Modern wireless technologies like orthogonal frequency-division multiplexing (OFDM), relaying, and spatial multiplexing allow high spectrum usage efficiency to be achieved, and some researchers argue that spectrum scarcity is a non-issue due to available technology [3]. Nonetheless, a bandwidth shortage of 275 MHz in the United States alone is foreseen by the end of 2014 [4].

To face these challenges, cellular operators are deploying complementary network infrastructure for data delivery, a technique known as mobile traffic offloading [5]. The two main technological advances to enable mobile traffic offloading are the introduction of small cell networks and the development of dynamic spectrum access techniques for operation in license-exempt radio bands.

The concept of small cells, as proposed for heterogeneous networks (HetNets), is two-fold. In the data plane, the goal is enabling the dense deployment of cells with smaller coverage areas, but capable of serving high traffic loads. On the other hand, in the control plane, the main goal is diminishing the dependence on an operator's backbone by implementing concepts like self-organization and self-adaptation. These requirements led 3GPP to standardize LTE small cells for operation on licensed spectrum in Release 12. 3GPP also foresees the adoption of enhanced IEEE 802.11 WLANs in unlicensed spectrum as a complementary solution. In this sense, IEEE 802.11ac networks with *Wi-Fi Passpoint* are a good starting point, while the IEEE 802.11ax standard (currently under development) is being considered for dense deployment scenarios. It is foreseen that by 2016, up to 30 percent of broadband access in cellular networks will be attained over traffic offloading networks [1].

Fuad M. Abinader, Jr., Erika P. L. Almeida, Fabiano S. Chaves, André M. Cavalcante, Robson D. Vieira, Rafael C. D. Paiva, and Angilberto M. Sobrinho are with Nokia Institute of Technology.

Sayantan Choudhury, Esa Tuomaala, and Klaus Doppler are with Nokia Research Center.

Vicente A. Sousa Jr. is with Federal University of Rio Grande do Norte.

0163-6804/14/$25.00 © 2014 IEEE

Dynamic spectrum access (DSA) has emerged as an alternative to overcome the increasing demand for additional capacity in wireless networks and spectrum scarcity [6]. DSA enablers like cognitive radio concepts motivated regulatory agencies to allow license-exempt operation in licensed spectrum. For instance, the United States [7] and Europe [8] recently published rules for operation of secondary users (SUs) in the so-called TV white spaces (TVWS). Another initiative is authorized shared access (ASA) [9], where incumbent spectrum holders negotiate their spectrum with SUs in underutilized locations while maintaining acceptable interference levels.

Despite small cells and DSA, spectrum demand is so intense that joint operation of LTE and Wi-Fi in the same license-exempt bands may be expected [10]. Although current spectrum allocation does not comprise any overlapped spectrum band between both technologies, there have been recent discussions in 3GPP concerning the need for feasibility studies about the deployment of LTE in unlicensed spectrum [11]. The objective of this study is to determine which enhancements would be needed from LTE to fulfill regulatory requirements to occupy those bands, for example, the 5.8 GHz industrial, scientific, and medical (ISM) band. Figure 1 shows the LTE and Wi-Fi spectrum allocation in the United States, already considering the 5.8 GHz ISM as a possibility for LTE deployment.

Coexistence of LTE and Wi-Fi in the same band poses certain technical challenges, and some performance degradation can be expected. From the early coexistence results in [12, 13], a series of questions could be made to direct future research: What issues arise from simultaneous operation of LTE and Wi-Fi in the same spectrum bands? What technology is affected the most? What can be done to improve performance of both networks while coexisting? Should enhancements be introduced for the physical (PHY) and/or media access control (MAC) layers?

This article discusses the coexistence of LTE and Wi-Fi networks in the same unlicensed spectrum bands from the radio resource management (RRM) perspective. We review the differences between LTE and Wi-Fi channel access mechanisms, and present recent results demonstrating the performance of the two technologies when they coexist in the same unlicensed band. Then we discuss coexistence mechanisms, including the adaptation of features in both LTE and Wi-Fi that may act as enablers for a coexistence scenario. Finally, we present future research trends and conclusions.

Challenges for LTE/Wi-Fi Coexistence in Unlicensed Bands

Enabling different networks to operate in the same shared spectrum requires taking some issues under consideration. One important aspect is coexistence, which involves the definition of boundaries for the occupation of radio resources (i.e. time and spectrum) by the networks, as well as intelligent modifications on the RRM algorithms to take into account coexisting dissimilar access technologies. Another important aspect is interworking, that is, intelligent management of user allocations among dissimilar access technologies, handling ongoing (e.g., handover) and incoming (e.g., access selection) connections. This work focuses on the coexistence aspects. Interworking and network selection are beyond the scope of this article.

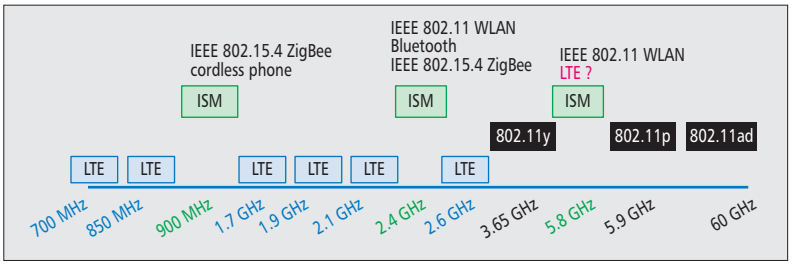


Figure 1. LTE and Wi-Fi spectrum allocation in the United States, considering future LTE deployment at the 5.8 GHz ISM band.

The lack of inter-technology coordination and mutual interference management are some of the main challenges for the efficient coexistence of different wireless technologies. Most broadband wireless access systems have interference management mechanisms, but these are designed to work properly for terminals of the same technology. These built-in mechanisms become less effective in heterogeneous wireless protocols/standards, which adopt asynchronous time slots, different channel access mechanisms, and disparate transmission/interference ranges. In fact, two of the most utilized broadband wireless access networks nowadays, LTE and Wi-Fi, are not only dissimilar but also incompatible when operating in the same band.

Wi-Fi employs OFDM for encoding digital data on multiple carrier frequencies, grouped within subcarriers where OFDM symbols are actually transmitted. In Wi-Fi infrastructure mode, an access point (AP) bridges a basic subscriber set (BSS) of wireless stations (STAs) to a wired Ethernet network. STAs and APs utilize a Wi-Fi default channel access mechanism, the distributed coordination function (DCF), to exchange data, control, and management frames. DCF uses a contention-based protocol known as carrier sense multiple access with collision avoidance (CSMA/CA), where nodes listen to the channel prior to transmission in a procedure known as clear channel assessment (CCA). A node in CCA may receive transmissions coming from other nodes, causing the channel to be seen as occupied, and hence deferring transmission to a random backoff time. CCA and backoff decrease the probability of transmission collisions in Wi-Fi at the cost of lower channel utilization.

On the other hand, LTE employs orthogonal frequency-division multiple access (OFDMA), which is a multi-user version of OFDM. Multiple access is achieved in LTE by assigning subsets of subcarriers to individual user equipments (UEs) for a specific number of symbol times (i.e., physical resource block, PRB), thus allowing simultaneous transmissions from several UEs. In

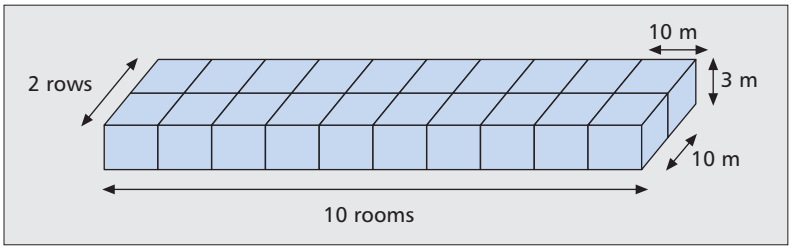


Figure 2. Single-floor/multi-room indoor scenario composed of 2 rows of 10 rooms, each measuring 10 m × 10 m × 3 m.

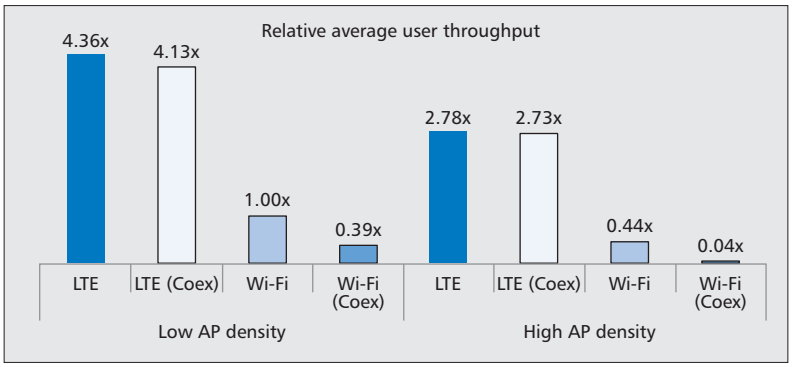


Figure 3. LTE and Wi-Fi average user throughput relative to Wi-Fi low AP density for indoor scenario. Deployments: low AP density (4 APs per technology) and high AP density (10 APs per technology) with an average STA density of 2.5 per AP for both cases. LTE and Wi-Fi evaluations: isolated (LTE, Wi-Fi) and in coexistence (LTE (Coex) and Wi-Fi (Coex)).

comparison with Wi-Fi using DCF, LTE has much more flexibility regarding resource allocation in time and frequency domains. Also, LTE does not require carrier sensing prior to transmission. Instead, the LTE base station (known as eNodeB) allocates radio communication subchannels for channel estimation and equalization, synchronization, management, control, and data transmissions. Finally, eNodeB deployment is usually planned, and inter-eNodeB communication infrastructure may be used for spectrum usage coordination.

Another challenge is the LTE deployment model for unlicensed spectrum bands. The first limiting factor is that regulatory agencies restrict the effective isotropic radiated power (EIRP) in unlicensed spectrum bands to much lower levels than typically used in LTE macrocells. Additionally, LTE should be able to determine whether Wi-Fi is jointly operating in the same spectrum as well as establishing a coexistence mechanism with it. From this, LTE small cells appear as a natural deployment model for LTE operation in unlicensed spectrum.

A potential traffic offloading scenario with coexisting LTE and Wi-Fi deployments is the single floor/multi-room indoor environment with LTE small cells and Wi-Fi, illustrated in Fig. 2. This scenario, composed of 2 rows of 10 rooms, each measuring 10 m × 10 m × 3 m, is adopted by both 3GPP and IEEE as a realistic scenario to represent residential and small office uncoordinated deployments.

In standalone single-floor/multi-room indoor deployments, it can be expected that LTE outperforms Wi-Fi in terms of average user throughput due to its more efficient usage of radio resources. A recent performance study [12] not only confirmed that, but also showed that when nodes of the two technologies coexist in the same frequency band, LTE interference severely affects Wi-Fi operation (Fig. 3).[1] The main reason is that LTE, in contrast to Wi-Fi, does not sense for channel vacancy prior to transmissions; thus, Wi-Fi nodes have a tendency to be blocked by LTE transmissions. Hence, while LTE is seldom affected by Wi-Fi interference, Wi-Fi is almost silenced when coexisting with LTE. This can be clearly seen on the average user throughput performance of both technologies presented in Fig. 3, especially for the high node density case.

The next section explores enabling features for efficient coexistence between LTE and Wi-Fi.

LTE/Wi-Fi Coexistence Enablers

Coexistence mechanisms can be broadly classified into collaborative and non-collaborative (autonomous), according to the exchange of messages between coexisting systems. Non-collaborative mechanisms may be used autonomously to facilitate coexistence with other networks and devices, while collaborative mechanisms require mutual agreement on the parameters used in each network.

A classic example of a non-collaborative coexistence mechanism is CSMA/CA with CCA in Wi-Fi, which enables coexistence with other wireless network technologies in unlicensed bands such as IEEE 802.15.4 (Bluetooth). On the other hand, a representative example of a standardized collaborative coexistence mechanism is IEEE 802.19.1, which defines a series of network elements, functions, and interfaces for the coexistence and coordination of different networks in TVWS bands. Utilization of collaborative coexistence mechanisms has greater potential to provide better performance for all coexisting networks than non-collaborative mechanisms. We describe a generalized procedure for collaborative coexistence through the flowchart in Fig. 4.

The generalized collaborative procedure assumes two operation modes: regular mode (RM) and coexistence mode (CM). RM represents standard operation, where no other technology is assumed to be using the spectrum at the same location and time. Here, the search for coexisting systems is done periodically, or triggered by external events such as the increase of the received interference or the detection of a beacon of another technology.

If a coexisting system is detected, the following actions are expected: identification of coexisting systems and synchronization with the identified systems. Synchronization can be done by reusing synchronization signals of the coexisting technology, such as the primary and secondary synchronization signals (PSS and SSS) of LTE and the preamble of Wi-Fi. Additional synchronization information can also be obtained by exploring the cyclic prefix repetitions of the coexisting technology.

Next, a negotiation phase is started. At this stage, the systems sharing the spectrum agree on

[1] Network performance assessed by system-level simulations modeling multi-cell and multi-user standard-compliant time-division duplexing LTE (TDD-LTE) and IEEE 802.11 (Wi-Fi). As detailed in [12, 14], these simulations include modeling of network layout, nodes distribution, radio environment, mutual interference, PHY and MAC layer, and traffic generation.

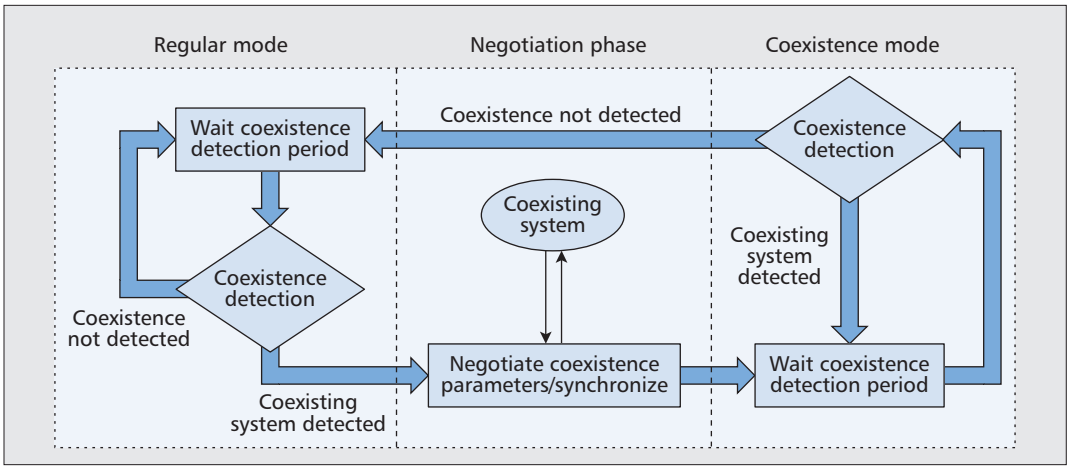


Figure 4. Generalized collaborative coexistence algorithm.

Non-collaborative mechanisms may be used autonomously to facilitate coexistence with other networks and devices, while collaborative mechanisms require mutual agreement on the parameters used in each network.

system parameters for a fair coexistence. Each system is expected to renounce some resources (e.g., time or frequency) it would use if operating in RM. However, minimum operational requirements of individual systems must be satisfied. If there are no mechanisms for communication between the coexisting technologies, each system should trigger coexistence techniques that avoid channel access domination by any of the coexisting radio access technologies.

Finally, each system reconfigures to agreed parameters, thus switching to CM. Once in CM, the systems monitor the shared resources in order to check whether there is effective operation of the coexisting systems. The system should also check for new secondary users, and return to the negotiation phase when necessary. When no coexisting system activity is detected, the operation is switched back to RM.

A number of *enabling features* can help with implementing collaborative coexistence mechanisms for LTE and Wi-Fi. For once, new spectrum utilization opportunities for both LTE and Wi-Fi can be created by allowing spectrum incumbents to grant access to subutilized licensed spectrum portions for secondary users, known as flexible spectrum access (FSA). In addition, given a specific spectrum portion, the selected *spectrum sharing technique* may operate on distinct dimensions (time, frequency, and space). Some LTE mechanisms for *interference management* can be adapted to enable coexistence with Wi-Fi STAs. On the other hand, since LTE interference has the potential to block Wi-Fi STAs using DCF, some features can improve Wi-Fi performance in coexistence with LTE, such as *channel selection* and *contention-free operation*. Table 1 presents a brief taxonomy of these enabling mechanisms for LTE/Wi-Fi coexistence according to the radio access technology.

Mechanisms listed in Table 1 can be used in CM operation within the general collaborative coexistence procedure illustrated in Fig. 4. Such coexistence enablers are described below.

Enabling mechanism	Technology
Flexible spectrum access	LTE/Wi-Fi
Channel selection	Wi-Fi
Blank subframes	LTE
Transmit power control	LTE

Table 1. Taxonomy for LTE/Wi-Fi coexistence enablers.

Flexible Spectrum Access

Two opposite spectrum licensing models regulate the operation of wireless broadband access systems. In the licensed model, a spectrum incumbent acquires exclusive utilization rights from governmental regulatory agencies via, for example, auctions. On the unlicensed model, spectrum portions are specifically allocated for non-exclusive utilization, and specific rules are set to ensure coexistence. DSA techniques have been considered as alternatives for improving utilization of spectrum portions. However, DSA efficiency in ensuring primary user rights, as well as the economic viability of systems operating under unpredictable portions of the spectrum, are aspects still under evaluation.

An alternative spectrum allocation model, FSA, has been recently proposed in the literature. One major example of FSA is authorized shared access (ASA) [9], where the spectrum incumbent economically explores its underutilized spectrum assets by granting exclusive access rights to third-party small cell systems, which also brings a number of complementary benefits. FSA grants can be constrained in frequency, time, and space, which ensures protection of incumbents and increases predictability for long-term investments in both the spectrum incumbent and the leasing third-party system. In addition, since small cells have a diminished coverage radius, they can be located nearer to the spectrum incumbent than conventional macrocells. The ASA approach takes advantage of existing products and standards such as LTE systems. As such, ASA has potential for being cost effective by reusing the available LTE infrastructure for complementary wireless service. Also, due to the opportunity for wider spectrum

The uncoordinated nature of Wi-Fi deployments and the limitation of non-overlapping channels in the ISM bands have motivated several studies about channel selection for Wi-Fi networks, which could also be exploited in coexistence with LTE.

	Subframe number									
Coexistence time	0	1	2	3	4	5	6	7	8	9
0 ms	D	S	U	U	D	D	S	U	U	D
2 × 1 ms	D	S	C	U	D	D	S	U	U	C
4 ms	D	C	C	C	C	D	S	U	U	D
2 × 2 ms	D	S	C	C	U	U	S	C	C	D

Table 2. Examples of null-subframe allocation considering LTE TDD. The coexistence time denotes the amount of time given for coexistence with Wi-Fi. D and U denote regular DL and UL subframes, respectively. S denotes a special subframe for signaling. C denotes coexistence subframes (with no LTE transmissions).

aggregation, it enables further performance improvements.

Channel Selection

Two major differences between Wi-Fi and LTE technologies are channel allocation and network deployment. While Wi-Fi was developed to be used in unlicensed bands with uncoordinated deployments, LTE was meant to be used in licensed spectrum bands and planned deployments. When both of them share the same frequency band, Wi-Fi performance is severely degraded by LTE transmissions, as discussed previously. Therefore, channel selection seems to be an important enabler for LTE/Wi-Fi coexistence.

The uncoordinated nature of Wi-Fi deployments and the limitation of non-overlapping channels in the ISM bands have motivated several studies about channel selection for Wi-Fi networks, which could also be exploited in coexistence with LTE. Some Wi-Fi access points (APs), for example, already implement simple channel selection techniques, such as least congested channel search (LCCS), in which the AP monitors its own channel, also searching for incoming packets from other APs, and selects the least congested one.

As a refinement, the subcarrier allocation flexibility provided by OFDM and OFDMA techniques can also be exploited in coexistence scenarios. Instead of fixed bandwidth channels, adaptive bandwidth channels could be defined and selected in coexistence scenarios.

Since Wi-Fi can be blocked by LTE when coexisting, it is in Wi-Fi's best interest to select the least congested channel for operation. In this case, some coordination between Wi-Fi APs and LTE eNodeBs for channel selection could ease the task of channel selection. This is an issue since exchange of information between nodes experiencing interference relies on a common intertechnology communication framework, which is currently unavailable for LTE and Wi-Fi.

Blank Subframes

An intuitive way to share the spectrum is avoiding technologies to access the channel at the same time. According to the discussion above, the probability of Wi-Fi being blocked when coexisting with LTE is high, and since regulatory rules usually mandate that technologies should share channel access in unlicensed spectrum, a time-sharing coexistence technique would require LTE silent periods. For this, a key LTE feature introduced in Release 10, the almost blank subframe (ABS), can be exploited. ABSs are LTE subframes with reduced downlink transmission power or activity, intended to coordinate transmission of macro and pico eNodeBs in heterogeneous deployments. During an ABS, LTE macro eNodeBs cause less interference to pico eNodeBs.

A modified version of ABS, where uplink (UL) and/or downlink (DL) subframes can be silenced, and no LTE common reference signals are included, is proposed in [14] as null-subframes to support coexistence with Wi-Fi. It is shown that Wi-Fi is able to reuse the blank subframes ceded by LTE, and that throughput increases with the number of null-subframes. Table 2 shows an example of null-subframe allocation inside an LTE frame. However, since LTE throughput decreases almost proportionally to the number of ceded blank subframes, a trade-off is established. Additional LTE performance degradation may be observed if blank subframes are nonadjacent, since Wi-Fi transmissions are not completely confined within LTE silent periods. This is illustrated in Fig. 5a, which summarizes the main results in [14]. However, if the duration and occurrence of LTE blank subframes is reported to Wi-Fi during the negotiation phase (Fig. 4), Wi-Fi nodes might be able to conveniently confine their transmissions within blank subframes and thus avoid interfering with LTE.

Transmit Power Control

LTE UL transmit power control is an alternative to the LTE blank subframes time-sharing approach for LTE/Wi-Fi coexistence. Here, a controlled decrease of LTE UEs' transmit powers diminishes the interference caused to neighboring Wi-Fi nodes, thus creating Wi-Fi transmission opportunities as Wi-Fi nodes detect the channel as vacant. Conventional LTE UL power control compensates only a fraction of the path loss. This effectively reduces LTE intercell interference, as UEs experiencing high path loss, usually in the cell edge, have their UL transmit power diminished. However, LTE power control

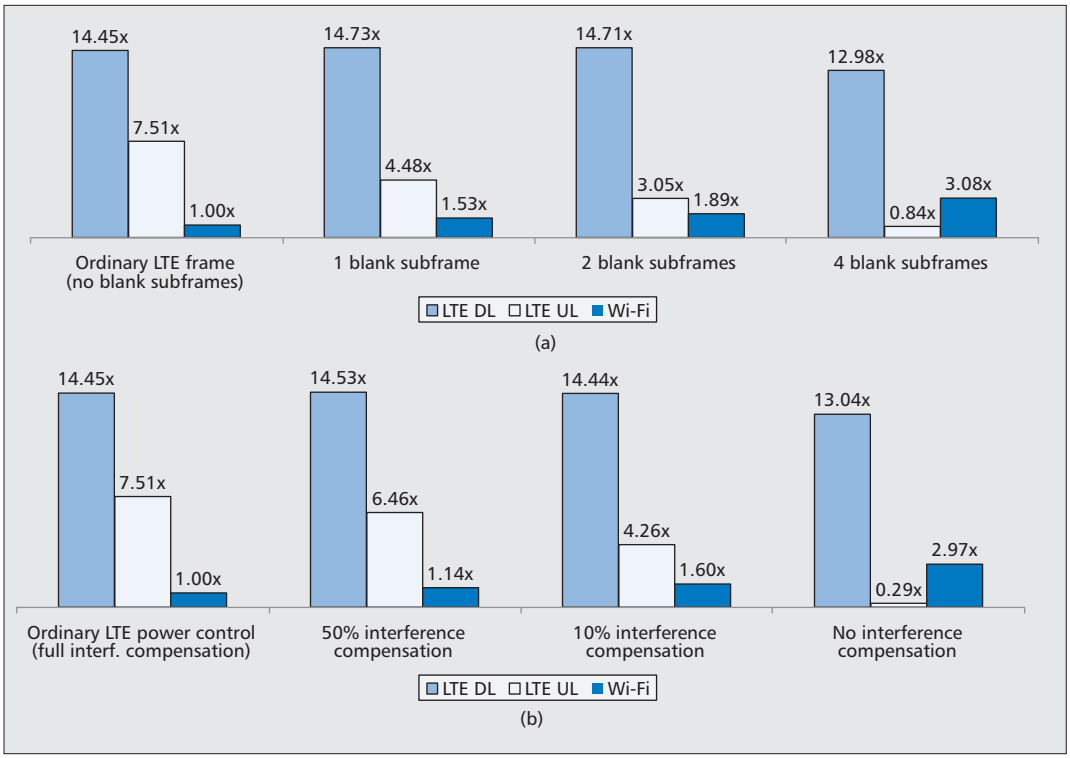


Figure 5. LTE and Wi-Fi average user throughput performance in coexistence with a deployment of 10 APs/25 STAs per technology in a 20-room single-floor indoor scenario, relative to Wi-Fi with no blank subframe. a) blank subframes allocation; b) LTE UL power control with an interference-aware operating point.

based on path loss is not effective for Wi-Fi coexistence, which requires transmit power reduction of UEs causing high interference to Wi-Fi nodes.

An LTE UL power control with an interference-aware power operating point is proposed in [15] for enabling coexistence with Wi-Fi. Interference measurements performed at LTE eNodeBs and/or UEs allow estimating the presence and proximity of Wi-Fi nodes. UEs measuring high interference are more likely to cause high interference, so UL transmit power is reduced according to a fractional compensation of the measured interference. LTE UL power control defines UE transmit powers so that path loss and interference are compensated and a given signal quality (i.e., a target signal-to-interference-plus-noise ratio, SINR), is achieved at the LTE eNodeB receiver. This fractional compensation of the measured interference corresponds to decreasing the target SINR when high interference is observed. As such, LTE UE throughput is decreased accordingly. As seen in Fig. 5b, the reduction of key LTE UEs' transmit powers indeed allows neighboring Wi-Fi transmissions at the cost of decreasing LTE throughput.

Simulated throughput results in Fig. 5 actually demonstrate that LTE blank subframes and UL transmit power control define different trade-off configurations for LTE and Wi-Fi in coexistence. While Fig. 5a shows the simultaneous Wi-Fi throughput increase and LTE throughput decrease with the number of blank subframes allocated, in Fig. 5b the decrease in the fraction of interference compensated by LTE UL transmit power control also decreases LTE throughput in favor of Wi-Fi throughput increase.

LTE/Wi-Fi Standardization for Coexistence in Unlicensed Bands

Studies undertaken so far clearly reveal that some challenges need to be addressed for the coexistence of LTE and Wi-Fi in the same unlicensed band. Standardization bodies (i.e., 3GPP and IEEE) are addressing some of these challenges.

From the perspective of LTE, 3GPP has recently started discussion on operation in unlicensed bands. A Study Item was created for defining modifications necessary to LTE radio for deployment in unlicensed spectrum [11]. On the other hand, with Wi-Fi conventionally operating in unlicensed bands, IEEE has worked on standardized mechanisms for allowing efficient coexistence among heterogeneous wireless broadband access technologies within TVWS bands. One major example of IEEE initiatives for operation in TVWS is the IEEE 802.19 Working Group (WG) [16], where a task group named 802.19 TG1 addresses coexistence for IEEE 802 networks and devices. These studies can also be useful for non-IEEE 802 networks and TV band devices (TVBDs). Another initiative is the IEEE 802.11af standard, published in February 2014, which covers Wi-Fi operation in the VHF and UHF bands between 54 and 790 MHz.

Studies undertaken so far clearly reveal that some challenges need to be addressed for the coexistence of LTE and Wi-Fi in the same unlicensed band. Standardization bodies (i.e., 3GPP and IEEE) are addressing some of these challenges.

With the increasing relevance of Wi-Fi for traffic offloading in cellular networks, improving Wi-Fi efficiency in terms of end-user performance in the presence of dense deployment of APs and STAs has become more important.

With the increasing relevance of Wi-Fi for traffic offloading in cellular networks, improving Wi-Fi efficiency in terms of end-user performance in the presence of dense deployment of APs and STAs has become more important. Recognizing this, IEEE 802 WG created the IEEE 802.11 High Efficiency WLAN (HEW) Study Group (SG) [2] in May 2013, aiming to enhance the quality of experience (QoE) of wireless users in everyday high-density scenarios. As a result of the discussions in HEW SG, the IEEE 802.11ax Task Group (TGax) was recently established to substantially increase user throughput in dense networks with a large number of users and devices, dense heterogeneous networks, and outdoor deployments. TGax includes improvements to cellular offloading as one of its major requirements, and is also investigating mechanisms to increase spatial capacity with PHY-MAC enhancements to the existing IEEE 802.11 standard in the 2.4 GHz and 5 GHz radio frequency bands. A first draft of TGax amendments to IEEE 802.11 is expected to be concluded by 2016.

Conclusions

The wireless communications community has been searching for solutions to handle the increasing demand for wireless broadband access. In this context of spectrum scarcity, there has been recent discussion about allowing wireless network technologies like LTE and Wi-Fi to coexist in the same unlicensed bands. In this article, we show that Wi-Fi is severely affected by concurrent operation of LTE in the same band. This indicates a serious need for coexistence mechanisms to improve the performance of both systems. The applicability of some coexistence enabling features for both LTE and Wi-Fi are discussed, and research directions for further development of inter-technology coexistence are presented. We also propose coexistence mechanisms by reusing the blank subframe approach and the UL transmit power used in LTE, and show that it can significantly improve Wi-Fi performance when coexisting with LTE in the same unlicensed bands.

References

[1] Cisco White Paper, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016," 2012.

[2] O. Aboul-Magd, IEEE 802.11 HEW SG Proposed Project Authorization Request (PAR), IEEE 802 WG Std. IEEE 802.11-14/0165r1; https://mentor.ieee.org/802.11/dcn/14/11-14-0165-01-0hew-802-11-hew-sg-proposed-par.docx

[3] G. Staple and K. Werbach, "The End of Spectrum Scarcity," *IEEE Spectrum*, vol. 41, no. 3, Mar. 2004, pp. 48–52.

[4] Deloitte, "Airwave Overload? Addressing Spectrum Strategy Issues that Jeopardize U.S. Mobile Broadband Leadership," Deloitte Development LLC, White Paper, Sept. 2012.

[5] C. Sankaran, "Data Offloading Techniques in 3GPP Rel-10 Networks: A Tutorial," *IEEE Commun. Mag.*, vol. 50, no. 6, 2012, pp. 46–53.

[6] I. F. Akyildiz *et al.*, "NeXt Generation/Dynamic Spectrum Access/Cognitive Radio Wireless Networks: A Survey," *Computer Networks*, vol. 50, no. 13, Sep. 2006, pp. 2127–59.

[7] "FCC 10-198 Notice of Inquiry," Nov. 2010, ET Docket no. 10-237.

[8] ECC, "Technical and Operational Requirements for the Operation of White Space Devices under Geo-Location Approach," Report 186, Jan. 2013.

[9] M. Matinmikko *et al.*, "Cognitive Radio Trial Environment: First Live Authorized Shared Access-Based Spectrum-Sharing Demonstration," *IEEE Vehic. Tech. Mag.*, vol. 8, no. 3, Sept. 2013, pp. 30–37.

[10] M. I. Rahman *et al.*, "License-Exempt LTE Systems for Secondary Spectrum Usage: Scenarios and First Assessment," IEEE Symp. New Frontiers in Dynamic Spectrum Access Networks, DySPAN, 2011, pp. 349–58.

[11] Q. Ericsson, *Study on LTE Evolution for Unlicensed Spectrum Deployments*, 3GPP TSG RAN Meeting 62, 3GPP TSG RAN Std. RP-131 788, Dec. 2013; http://www.3gpp.org/ftp/tsg ran/TSG RAN/TSGR 62/Docs/RP-131788.zip

[12] A. M. Cavalcante *et al.*, "Performance Evaluation of LTE and Wi-Fi Coexistence in Unlicensed Bands," *Proc. IEEE 77th VTC 2013-Spring*, Dresden, Germany, June 2013.

[13] T. Nihtil *et al.*, "System Performance of LTE and IEEE 802.11 Coexisting on a Shared Frequency Band," *IEEE Wireless Commun. and Networking Conf. 2013*, Apr. 2013.

[14] E. P. L. Almeida *et al.*, "Enabling LTE/Wi-Fi Coexistence by LTE Blank Subframe Allocation," *Proc. IEEE ICC '13*, 2013.

[15] F. S. Chaves *et al.*, "LTE UL Power Control for the Improvement of LTE/Wi-Fi Coexistence," *Proc. IEEE VTC 2013-Fall*, Las Vegas, NV, Sept. 2013.

[16] T. Baykas, M. Kasslin, and S. Shellhammer, "IEEE 802.19.1 System Design Document," IEEE 802 WG, Mar. 2010.

Biographies

Fuad M. Abinader, Jr. (fuad.abinader@indt.org.br) received his B.Sc. in computer science from Federal University of Amazonas (UFAM) in 2003, and his M.Sc. in informatics from UFAM in 2006. He is currently a doctoral student in electrical engineering at Federal University of Rio Grande do Norte (UFRN) and a researcher at Nokia Institute of Technology (INdT). His current interests include mobile Internet protocols, Wi-Fi, LTE, and WiMAX, and standardization in IEEE, 3GPP, and IETF.

Erika P. L. Almeida (erika.almeida@indt.org.br) received her B.Sc. in telecommunications engineering and M.Sc. degrees from the University of Brasilia (UnB), Brazil, in 2007 and 2010, respectively. She has been a researcher at INdT since 2011, where she has worked on LTE, white space concepts and coexistence issues in TV white spaces. Her current research topics include Wi-Fi evolution and cognitive radio networks.

Fabiano S. Chaves (fabiano.chaves@indt.org.br) received his B.Sc. and M.Sc. degrees in electrical engineering from the Federal University of Ceará (UFC), Brazil, and his Ph.D. degree in electrical engineering from the University of Campinas (UNICAMP), Brazil, and the École Normale Supérieure de Cachan, France, in 2010. Since 2010, he has been a research engineer at INdT, Brazil. His research interests include radio resource management, signal processing and game theory for communications, and cognitive radio systems.

André M. Cavalcante (andre.cavalcante@indt.org.br) received his B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Federal University of Pará (UFPA) in 2001, 2003, and 2007, respectively. Since 2007 he has worked as a research engineer at INdT on several research projects and standardization activities (IEEE). His areas of interest are the evolution of Wi-Fi networks, beyond 4G networks, and systems with multiple antennas.

Robson D. Vieira (robson.d.vieira@indt.org.br) received M.Sc. and Ph.D. degrees in electrical engineering from the Catholic University of Rio de Janeiro, Brazil, in 2001 and 2005, respectively. From 2005 to 2010, he worked with white space concepts, and supporting some GERAN and 802.16m standardization activities focused on system performance evaluation at INdT. Since 2010, he is an R&D technical manager at INdT. His research interests include Wi-Fi Evolution, B4G, and cognitive radio networks.

Rafael C. D. Paiva (rcdpaiva@yahoo.com.br) has been a researcher at INdT since 2008. He obtained his Bachelor's degree in electrical engineering from Federal University of Santa Maria (UFSM) in 2005, his Master's degree in signal

processing from the Federal University of Rio de Janeiro (UFRJ) in 2008, and his Doctor's degree in acoustics and audio signal processing from Aalto University in 2013. Among his research interests are digital signal processing and new technologies of wireless networks.

ANGILBERTO M. SOBRINHO (angilberto.m.sobrinho@indt.org.br) received his M.Sc. degree in computer architectures from the Industrial Engineering Faculty (FEI) in 1984, and his M.Sc. in industrial automation from Federal University of Campina Grande (UFCG) in 2006. He joined as a professor of the State University of Amazonas (UEA) in 2005, and since 2012 has been working as a researcher at INdT. His main areas of interest are time synchronization in packet networks, and adaptive and phased array antennas.

SAYANTAN CHOUDHURY (sayantan.choudhury@nokia.com) leads the wireless research and standardization activities in NRC-Berkeley. His interests include optimization of PHY and MAC layers focusing on LTE-Advanced and next generation Wi-Fi networks. Currently, he is investigating concepts to enable dense deployments of Wi-Fi and also coexistence of LTE and Wi-Fi systems in unlicensed bands. He is the co-recipient of the 2009-2010 Sharp Labs Inventor of the Year award, and the *2010 IEEE Transactions on Multimedia* and 2012 PIMRC Best Paper awards.

ESA TUOMAALA (esa.tuomaala@nokia.com) received his M.S.(Tech.) degree in engineering physics and mathematics from Helsinki University of Technology, Espoo, Finland, in 2002. He joined Nokia Research Center in 2000. He is currently working as a principal researcher, focusing on system-level performance evaluation of next generation wireless systems and contributing to the development of relevant IEEE standards.

KLAUS DOPPLER (klaus.doppler@nokia.com) received his Ph.D. from Helsinki University of Technology, Finland, in 2010 and his M.Sc. in electrical engineering from Graz University of Technology, Austria, in 2003. He joined Nokia Research Center in 2002 and currently leads the Wireless Systems team in Berkeley, California. He has been recognized several times as top inventor in Nokia. He has about 75 pending and granted patent applications, and published in 30 journals, conference publications, and book chapters.

VICENTE A. DE SOUSA, JR. (vicente.sousa@ct.ufrn.br) received his B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from UFC in 2001, 2002, and 2009, respectively. Between 2001 and 2006, he developed solutions to UMTS/WLAN interworking for UFC and Ericsson of Brazil. Between 2006 and 2010, he contributed to WIMAX standardization and Nokia's product as a researcher at INdT. He is now a lecturer at UFRN, Brazil.

5G NETWORKS: END-TO-END ARCHITECTURE AND INFRASTRUCTURE




David Soldani *Kostas Pentikousis* *Rahim Tafazolli* *Daniele Franceschini*

The forthcoming 5G infrastructure, defined as the ubiquitous ultra-broadband network enabling the future Internet (FI), is not only about new releases of current network generations and services, but, more significantly, will be associated with a true revolution in the information and communications technologies (ICT) field: the network will efficiently and effectively take forward new-fangled services to everyone and everything, such as cognitive objects and cyber physical systems (CPSs). A "full immersive (3D) experience" enriched by "context information" and, in particular, "anything or everything as a service (XaaS)" are the main business drivers for massive adoption and market uptake of the new fundamental enabling technologies, beyond today's "client-server" model, where the network has been reduced to a ubiquitous "pipe of bits." XaaS refers to those services — beyond the current models of software as a service (SaaS), infrastructure as a service (IaaS), and platfore as a service (PaaS), SPI models, of cloud computing — such as data as a service (DaaS), security as a service (again, SaaS), network as a service (NaaS), knowledge as a service (KaaS), machine as a service (MaaS), and robot as a service (RaaS), which could be delivered over the advanced 5G infrastructure, without the need to own hardware, software, or even the cognitive objects themselves. Communication services, such as voice and video telephony, will be enriched and bundled with other services. The network infrastructure is expected to become the "nervous system" of the actual *digital society and digital economy*. This challenge calls for a complete redesign of services and service capabilities, architectures, interfaces, functions, access and non-access stratum protocols and related procedures, as well as advanced algorithms (e.g., for unified connection, security, mobility and routing management, and reconfiguration of ICT services; and any type of resource of cyber physical systems). The expected transformation will be especially true at the edge, that is, around the end user (or *prosumer*), where the "intelligence" already started migrating a few years ago, and where massive processing, memory, and storage capacity are gradually accumulating.

As of today, many challenges remain to be addressed to meet the expected key performance indicators and new services in terms of, for example:

- Throughput: 1000× more in aggregate and 10× more at link level
- Latency: 1 ms for remote control of robots or tactile Internet applications, and below 5 ms for 2–8K change in view at 30–50 Mb/s
- Ultra-high reliability
- Coverage suitable for a seamless experience
- Battery lifetime: 10× longer
- Spectrum utilization: all spectra, from cellular bands to visible light

The redesign of the radio access nodes will required innovation in multiple areas of basic radio technologies, such as a new air interface, new virtualized radio access networks, new radio frequency transceiver architecture, and new device radio architecture. New radio backhaul/fronthaul and new fiber access for the fixed network to support 5G wireless are also required as an integral part of the solution. The software defined 5G architecture running on a fully integrated wireless/optical infrastructure will be the de facto platform for future carrier networks on the horizon of 2020 and beyond. This *plastic* architecture will unify connection, security, mobility, and routing management, especially for supporting diverse vertical industry applications. Full compatibility with current and future incremental 4G releases will be guaranteed by the possibility of instantiating any type of virtual architecture and installing any kind of network and service application efficiently.

In *IEEE Communications Magazine*, this timely Feature Topic brings together key contributions of researchers from industry and academia that address the above challenging issues, and presents the fundamental peculiarities of the advanced 5G infrastructure to be open and flexible enough to meet defined (current) and unidentified (future) stakeholders' requirements. It also presents how the widespread adoption and utilization of cloud computing at the edge, software defined networking (SDN), services,

and network function virtualization (NFV) will make the 5G infrastructure technically feasible and, especially, business viable.

In response to our Call for Papers, 30 submissions were received. The submissions underwent a rigorous review process, following which only four outstanding contributions were selected for publication. The four articles are in the broad area of 5G network architectures, mobility and routing management, and device-to-device (D2D) communications. These articles are expected to stimulate new ideas and contributions within the research community, in addition to providing readers with relevant background information and feasible solutions to the main technical design issues of future 5G networks.

The first article, "Design Considerations for a 5G Network Architecture" is by Patrick Kwadwo Agyapong, Mikio Iwamura, Dirk Staehle, Wolfgang Kiess, and Anass Benjebbour. The authors present a two-layer 5G network architecture consisting of a radio access network and a network cloud, which integrates many enabling technologies such as small cells, massive MIMO, SDN, and NVF to facilitate optimal use of network resources for QoE provisioning and planning. In this article, an initial proof of concept is also presented in order to demonstrate the technical feasibility of the proposed architecture. The crucial issues that need to be addressed and resolved to realize a complete 5G architecture vision are also thoroughly discussed.

The second article, "A New Control Plane for 5G Network Architecture with a Case Study on Unified Handoff, Mobility, and Routing Management by Volkan Yazıcı, Ulaş C. Kozat, and M. Oğuz Sunay, proposes an all-SDN network architecture with hierarchical network control capabilities as a simplified and unified approach to mobility and routing management for 5G networks. Beyond this, the novel architecture supports connectivity management as a service (CMaaS), which may be offered to provide QoS differentiation with a range of options to protect flows against subscriber mobility at different price levels without the utilization of tunneling protocols. Performance results show the proposed architecture to be a viable solution for mobile D2X communications with end-to-end latency below 5 ms.

In the third article, "Terminal-Centric Distribution and Orchestration of IP Mobility for 5G Networks," Alper Yegin, Jungshin Park, Kisuk Kweon, and Jinsung Lee first describe the efficiency issues of the current Third Generation Partnership Project (3GPP) centralized approach to mobility management using core anchoring. Then they present and discuss multiple dimensions of distributing the mobility functions, and suggest how the mobile terminal can utilize them (terminal based solution) in orchestration for efficient communication over a 5G flat mobile network. An L4+ mobility solution is the preferred choice when a flow requires session continuity and both endpoints support at least one common L4+ mobility protocol that is applicable to the flow. Remote anchoring is preferred over access anchoring as the former does not create a triangular data path. Access anchoring acts as a supplement when an L4+ solution or core/remote anchoring is used, and provides the optimal data path between the terminal and its peer (remote end) only at the beginning of the communication, when the terminal is still close to the anchoring point.

The fourth article, "Toward D2D-Enhanced Heterogeneous Networks," by Francesco Malandrino, Claudio Casetti, and Carla Fabiana Chiasserini, argues the need to integrate functionally D2D and infrastructure-to-device (I2D) modes, and proposes a multi-modal proportional fairness (MMPF) algorithm to achieve this goal. They evaluate the impact of D2D in a two-tier scenario, where macro and micro coverage are combined. Simulation results show that although I2D retains a clear edge for general-purpose downloading, D2D is a viable solution for localized transfers as well as viral content. Ultimately, D2D can indeed be integrated in cellular networks (e.g., as far as the network control concerns for the optimized usage of the spectrum, and for authentication and authorization), but apparently cannot entirely replace the deployment of small cells: it is useful to complement them and mitigate the effects of a reduced infrastructure deployment together with enabling new services based on proximity in a scalable way.

Many initiatives on 5G are currently ongoing at the global level. For example, in the United States, the three main activities carried out on 5G are in the Intel Strategic Research Alliance (ISRA), 4G Americas, and NYU Wireless Research Center. In China, it is ongoing in the Ministry of Science and Technology (MOST) 863 Research Program (Chinese: 863计划) and IMT-2020 (5G) Promotion Group. In Japan, the 2020 and Beyond Ad Hoc Group is under the Association of Radio Industries and Businesses' (ARIB's) advanced wireless communications study committee. In Korea, the main activity is in the 5G Forum. The most important initiatives in the European Union are the 5G Private Public Partnership (5G PPP) and the 5G Innovation Centre (5G IC), at the University of Surrey, United Kingdom. The 5G PPP is within the EU Horizon 2020 — The EU Framework Programme for Research and Innovation — under one of the most important EU Industrial Leadership challenges: ICT-14 Advanced 5G Network Infrastructure. Within this research and innovation framework, the European Commission (EC), under the approval of the European Parliament (EP), has already committed €700 million of public funds over six years (2015–2021). From two to ten times higher is expected to be the investment from private parties: industry, SME, and research institutes.

Most efforts are currently focused on research and innovation work; after that, intensive standardization activities, field tests, and large-scale trials will take place to accelerate industrial pre-adoption. Commercial products will likely be available in the market beyond 2020. This approximate roadmap applies to infrastructures and devices for human and, in particular, mission-critical and massive machine communications.

In closing, we would like to thank all the stakeholders who have made this Feature Topic possible: colleagues who spread the Call for Papers around, the many authors who submitted papers to our Feature Topic, the team of reviewers, who helped us to select and further improve the outstanding papers that are published in this magazine, the Editor-in-Chief (EiC) of *IEEE Communications Magazine*,

Dr. Sean Moore, as well as the members of the Editorial Board for their invaluable help and for hosting this Feature Topic, and ComSoc's editorial staff who produced the final material.

We hope that this endeavor will meet readers' expectations, for whom this Feature Topic on advanced 5G infrastructure has been prepared.

BIOGRAPHIES

DAVID SOLDANI received an M.Sc. degree with maximum score and cum laude approbatur in electron ic engineering from the University of Florence, Italy, in December 1994, and a D.Sc. degree in technology with distinction from Aalto University, Finland, in October 2006. He is one of the top experts in multi-disciplinary, long-term, transformative frontier research and innovation. He has been active in the ICT field for more than 20 years, successfully working on more than 150 R&D projects for 2–5G, generating original contributions to all types of quality deliverables: from strategic research and innovation to modeling, simulations, emulations, and innovative proof of concepts with stakeholders. He is currently VP of the Huawei European Research Centre (ERC), Germany, and visiting professor at the University of Surrey, United Kingdom. Areas of his responsibility and expertise include, but are not limited to, future wireless, network, IoT, and multimedia technologies. He represents Huawei on the Board of the 5G Infrastructure Association and Steering Board of NetWorld2020 European Technology Platform in Europe.

KOSTAS PENTIKOUSIS is the head of IT Infrastructure at the European Center for Information and Communication Technologies (EICT GmbH), Berlin, Germany. He earned his Bachelor's degree in informatics (1996) from Aristotle University of Thessaloniki, Greece, and his Master's (2000) and doctoral (2004) degrees in computer science from Stony Brook University. His research interests include network architecture, system, and protocol design. Visit http://linkedin.com/in/kostas for more details.

RAHIM TAFAZOLLI is director of the Institute for Communication Systems (ICS) and the 5G Innovation Centre (5GIC), Faculty of Engineering, and Physical Sciences, University of Surrey. He has published more than 500 research papers in refereed journals and international conferences, and as an invited speaker. He is the editor of two books, *Technologies for Wireless Future* (Wiley: Vol. 1 2004, Vol. 2 2006). He is currently chairman of EU Net!Works Technology Platform Expert Group and a board member of the U.K. Future Internet Strategy Group (UK-FISG). He was appointed a Fellow of the Wireless World Research Forum in April 2011 in recognition of his personal contribution to the wireless world. He also heads one of Europe's leading research groups.

DANIELE FRANCESCHINI is currently responsible for technology, development plans, and cost analysis in the Network Planning Department, Latam. He joins other relevant departments in the definition of the group technology plan, technology network development plans, and the technical economical evaluation for cost analysis inside technology. Recently, from May 2011 to March 2013, he was responsible for the Next Generation Mobile Network strategy in Telecom Italia, in charge of all the group strategic aspects related to the Telecom Italia Mobile Broadband evolution and, in particular, the LTE introduction in Italy and in Latam. In 1998 he accompanied the UMTS standardization process, joining the ETSI SMG2 L2&3 group, and subsequently 3GPP TSG RAN WG2 and WG4, with editorship responsibilities of the specifications of the Radio Resource Management. He also worked on 3GPP SA1 and SA2. From March 2000 he joined Omnitel where he followed aspects related to UMTS. In 2001 he rejoined Telecom Italia Lab, where he worked on issues related to implementation of UTRAN, radio resource management, radio protocol architecture, and UMTS in general in Italy and for the Telecom Italia companies in Europe and in Latam. Since the end of 2005 he was responsible for wireless access innovation activities within Telecom Italia Lab with particular focus on all of the radio innovation aspects and technologies related to the radio access network and its evolution (e.g., HSPA evolution, LTE, and LTE Advanced). he holds a degree in telecommunication engineering from the University of Pisa and currently is part of NGMN as an alternate Board Director, and follows the ITU-R WP5D Spectrum Group and the GSMA SSMG (Spectrum Strategy Management Group) working group.

Design Considerations for a 5G Network Architecture

Patrick Kwadwo Agyapong, Mikio Iwamura, Dirk Staehle, Wolfgang Kiess, and Anass Benjebbour

ABSTRACT

This article presents an architecture vision to address the challenges placed on 5G mobile networks. A two-layer architecture is proposed, consisting of a radio network and a network cloud, integrating various enablers such as small cells, massive MIMO, control/user plane split, NFV, and SDN. Three main concepts are integrated: ultra-dense small cell deployments on licensed and unlicensed spectrum, under control/user plane split architecture, to address capacity and data rate challenges; NFV and SDN to provide flexible network deployment and operation; and intelligent use of network data to facilitate optimal use of network resources for QoE provisioning and planning. An initial proof of concept evaluation is presented to demonstrate the potential of the proposal. Finally, other issues that must be addressed to realize a complete 5G architecture vision are discussed.

INTRODUCTION

Despite the advances made in the design and evolution of fourth generation cellular networks, new requirements imposed by emerging communication needs necessitate a fifth generation (5G) mobile network. New use cases such as high-resolution video streaming, tactile Internet, road safety, remote monitoring, and real-time control place new requirements related to throughput, end-to-end (E2E) latency, reliability,[1] and robustness[2] on the network. In addition, services are envisioned to provide intermittent or always-on hyper connectivity for machine-type communications (MTC), covering diverse services such as connected cars, connected homes, moving robots, and sensors that must be supported in an efficient and scalable manner. Furthermore, several emerging trends such as wearable devices, full immersive experience (3D), and augmented reality are influencing the behavior of human end users and directly affecting the requirements placed on the network. At the same time, ultra-dense small cell deployments and new technologies such as massive multiple-input multiple-output (mMIMO), software defined networking (SDN), and network function virtualization (NFV) provide an impetus to rethink the fundamental design principles toward 5G.

This article proposes a novel 5G mobile network architecture that accommodates the evolution of communication types, end-user behavior, and technology. The article first highlights trends in end-user behavior and technology to motivate the challenges of 5G networks. Some potential enablers are identified, and design principles for a 5G network are highlighted. This is followed by the articulation of a 5G mobile network architecture together with details of some fundamental technology enablers and design choices, and a discussion of issues that must be addressed to realize the proposed architecture and an overall 5G network. The article wraps up with proof of concept evaluations and conclusions.

CURRENT TRENDS

It is well known that mobile data consumption is exploding, driven by increased penetration of smart devices (smartphones and tablets), better hardware (e.g., better screens), better user interface design, compelling services (e.g., video streaming), and the desire for anywhere, anytime high-speed connectivity. What is perhaps not widely mentioned is that more than 70 percent of this data consumption occurs indoors in homes, offices, malls, train stations, and other public places [1]. Furthermore, even though mobile data traffic is increasing at a brisk pace, signaling traffic is increasing 50 percent faster than data traffic [2].

More end users are using multiple devices with different capabilities to access a mix of best effort services (e.g., instant messaging and email) and services with quality of experience (QoE) expectations (e.g., voice and video streaming). Over-the-top (OTT) players provide services and apps, some of which compete directly with core operator services (e.g., voice, SMS, and MMS). Connectivity is increasingly evaluated by end users in terms of how well their apps work as expected, regardless of time or location (in a crowd or on a highway), and they tend to be unforgiving toward the mobile operator when these expectations are not met. Moreover, the battery life of devices and a seamless experience across multiple devices (or a device ecosystem)

Patrick Kwadwo Agyapong, Mikio Iwamura, Dirk Staehle, and Wolfgang Kiess are with DOCOMO Communications Laboratories Europe GmbH.

Anass Benjebbour is with NTT DOCOMO Inc.

[1] In [3], reliability is defined as "the probability that a certain amount of data to or from an end user device is successfully transmitted to another peer (e.g., Internet server, mobile device, sensor, etc.) within a predefined time frame, i.e., before a certain deadline expires. The amount of data to be transmitted and the deadline are dependent on the service characteristics."

[2] In the context of this article, robustness is defined as the ability of the network to support a minimum predefined service level (e.g., minimum signal-to-interference-plus-noise ratio, SINR, to support basic voice communications) regardless of the network conditions (e.g., in natural disasters).

0163-6804/14/$25.00

Based on current trends, it is generally understood that 5G mobile networks must address six challenges that are not adequately addressed by state-of-the-art deployed networks: higher capacity, higher data rate, lower E2E latency, massive device connectivity, reduced capital and operations cost, and consistent QoE provisioning.

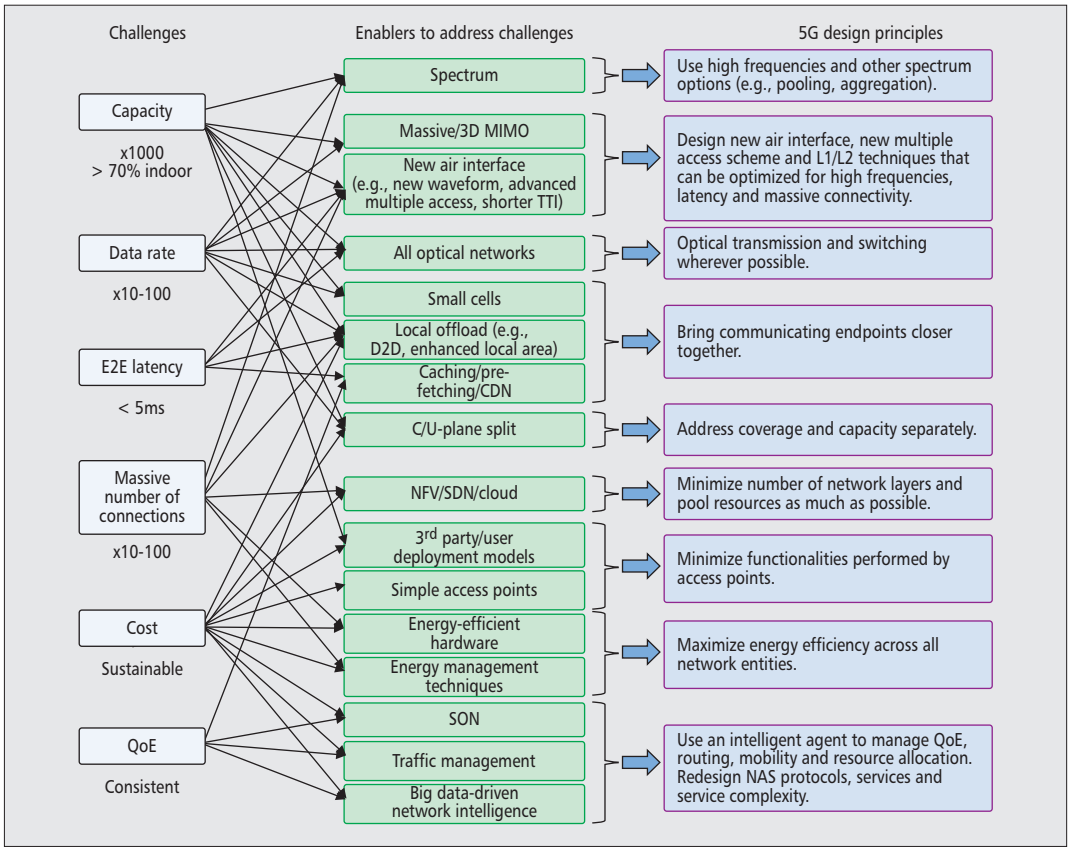


Figure 1. 5G challenges, potential enablers, and design principles.

have also become important issues for many end users.

The Internet of Things (IoT), which adds "anything" as an additional dimension to connectivity (in addition to anywhere and anytime), is also becoming a reality. Smart wearable devices (e.g., bracelets, watches, glasses), smart home appliances (e.g., televisions, fridges, thermostats), sensors, autonomous cars, and cognitive mobile objects (e.g., robots, drones) promise a hyper-connected smart world that could usher in many interesting opportunities in many sectors of life such as healthcare, agriculture, transportation, manufacturing, logistics, safety, education, and many more. Even though operators currently rely on existing networks (especially widely deployed 2G/3G networks and fixed line networks) to support current IoT needs, many of the envisaged applications impose requirements, such as, very low latency and high reliability, that are not easily supported by current networks.

To cope with such evolving demands, operators are continuously investing to enhance network capability and optimize its usage. Operators are deploying more localized capacity, in the form of small cells (e.g., pico and femto cells and remote radio units, RRUs, that are connected to centralized baseband units by optical fiber) to improve capacity. In addition, traffic offloading to fixed networks through local area technologies such as Wi-Fi in unlicensed frequency bands has become widespread. To optimize network usage for better QoE in a fair manner, mobile networks are also integrating more functionality such as deep packet inspection (DPI), caching, and transcoding. All these improvements come at significant capital and operating costs, however.

With the increasing complexity and associated costs, several concepts and technologies that have proved useful to the information technology (IT) sector are becoming relevant to cellular networks as well. For instance, an industry specification group (ISG) set up under the auspices of the European Telecommunications Standards Institute (ETSI ISG NFV) is currently working to define the requirements and architecture for the virtualization of network functions and address identified technical challenges. Similarly, the Open Networking Foundation approved a Wireless and Mobile Working Group in November 2013 to identify use cases in the wireless and mobile domain that can benefit from SDN based on OpenFlow.

5G Challenges, Enablers, and Design Principles

Based on current trends, it is generally understood that 5G mobile networks must address six challenges that are not adequately addressed by state-of-the-art deployed networks (Long Term Evolution-Advanced, LTE-A): higher capacity, higher data rate, lower E2E latency, massive device connectivity, reduced capital and operations cost, and consistent QoE provisioning [3, 4]. These challenges are briefly discussed below together with some potential enablers to address them. Figure 1 provides an overview of the challenges, enablers,[3] and corresponding design prin-

[3] The connections between the challenges and enablers depict the most significant linking, but not necessarily all possible connections.

ciples for 5G. It must be noted that the enablers highlighted in Fig. 1 also introduce their own set of challenges and corresponding key performance indicators (KPIs). Some of these challenges are discussed in the relevant sections. Nevertheless, a detailed discussion of the relevant KPIs is outside the scope of this article. The interested reader is referred to [4] for more details on this aspect.

System Capacity and Data Rate

Beyond 2020 mobile networks need to support a 1000-fold increase in traffic relative to 2010 levels, and a 10- to 100-fold increase in data rates even at high mobility and in crowded areas if current trends continue [1, 3, 4]. This requires not only more capacity in the radio access network (RAN), but equally important, also in the backbone, backhaul, and fronthaul. Pricing schemes can be used to manage and potentially reduce the increase in data consumption, as already demonstrated by operators in the market. However, as customers are willing to pay for the provisioned service rather than the data volume, pricing models may not be effective to suppress traffic in the future.

The current consensus is that a combination of more spectrum, higher spectrum efficiency, network densification, and offloading are necessary to address these challenges in the RAN [5]. Opportunities for more spectra include higher frequency bands (e.g., millimeter-wave, mmW), unlicensed spectrum, and aggregation of fragmented spectrum resources using carrier aggregation techniques. Dual connectivity of terminals to multiple base stations can exploit aggregated use of spectrum deployed at different base stations. Besides the available bandwidth, high frequency bands also allow for mMIMO using antenna arrays with small form factors, which can provide a 10-fold increase in capacity compared to conventional single-antenna systems [6]. Nevertheless, high frequency bands suffer from high path loss attenuation and are limited to line of sight (LOS) and short-range non-LOS environments. Massive MIMO can be exploited to extend the coverage of higher frequency bands by relying on beamforming gains.

Advanced physical layer techniques, such as higher-order modulation and coding schemes (MCS), such as 256-quadrature amplitude modulation (QAM), increase spectral efficiency and can be combined with mMIMO to increase system capacity. By adding some intelligence at the transmitter and receiver, potential interference can be coordinated and cancelled at the receiver to increase system throughput [7]. With such techniques in place, new schemes such as non-orthogonal multiple access (NOMA), filter bank multicarrier (FBMC), and sparse coded multiple access (SCMA) can further be utilized to improve spectral efficiency. For example, NOMA with successive interference cancelling (SIC) receivers has been shown to improve overall throughput in macrocells compared to orthogonal multiple access schemes by up to 30 percent even for high-speed terminals, with further gains expected with advanced power control [8].

Network densification refers to the dense deployment of many small cells. High carrier frequencies are well suited for small cells. The high attenuation they suffer is no longer seen as a drawback, but rather as an enabler to provide effective separation and mitigate interference between densely deployed small cells. To allow efficient improvement of capacity at critical locations, it is desirable that coverage and capacity be addressed independently. This can be realized through an architecture where control (C) and user data (U) planes are split among different cells [9]. The benefit of this approach is that U-plane resources can be scaled independent of C-plane resources. This allows more U-plane capacity to be provided in critical areas where it is needed, without the need to also provide co-located C-plane functionalities. Thus, more flexible deployments at lower costs can be realized. In such a C/U-plane split architecture, macrocells can provide coverage (C+U), and small cells can provide localized capacity (U).

Techniques like mMIMO and higher-order MCS can be employed in small cells to boost throughput [5]. Massive MIMO has an increased risk of link failure due to narrow beamforming, but this could be mitigated by employing robust techniques like dual connectivity, which always provides uninterrupted fallback to the coverage layer. Additionally, local offload through techniques such as network-controlled device-to-device (D2D) communications can further increase achievable system throughput [10].

Advances in optical networking, including optical switching, may be able to address the capacity requirements in the backbone, backhaul, and fronthaul. In addition, mMIMO can be used to provide high-capacity wireless backhaul and fronthaul links in LOS conditions.

End-to-End Latency

End-to-end latency is critical to enable new real-time applications. For example, remote controlled robots for medical, first response, and industrial applications require rapid feedback control cycles in order to function well. Safety-critical applications for cars and humans, built around vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication, also require very quick request-response and feedback control cycles with high availability and reliability. Augmented and virtual reality applications (e.g., immersive displays and environments) require very fast request-response cycles to mitigate cyber sickness. In order to realize these applications, networks must be able to support a target of 1 ms E2E latency with high reliability [11].

Innovations in air interface, hardware, protocol stack, backbone, and backhaul (all-optical transmission and switching), as well as network architecture can all help to meet this challenge. A new air interface with new numerology, such as shorter transmission time interval (TTI), can reduce over-the-air latency to a few hundred microseconds. Shorter TTI requires high available bandwidth, but this can be supported by using higher frequency bands. Note that such new numerology relies on significant improvements in receiver hardware (e.g., processing power and buffer size).

In addition, E2E latency can be reduced by enhancements in higher-layer protocols (e.g., use case and network-aware admission/congestion control algorithms to replace TCP slow start),

To allow efficient improvement of capacity at critical locations, it is desirable that coverage and capacity be addressed independently. This can be realized through an architecture where control (C) and user data (U) planes are split among different cells.

Pooled hardware resources can be shared among multiple functions, thus realizing multiplexing gains and lowering the amount of necessary hardware. The flexibilities enabled by NFV and SDN can make the network quick to deploy and more adaptable, and reduce time to market for new services.

bringing communicating endpoints closer (e.g., through network-controlled D2D and ultra-dense small cell deployments with local breakout) and adding more intelligence at the edge of the network. The latter is realized, e.g., through caching and pre-fetching techniques, service-dependent location of C-plane protocols and orchestration. For example, C-plane protocols necessary for latency-critical MTC services may be distributed at the edge of the network, whereas C-plane protocols required for services with more relaxed latency requirements could be located at a central entity. Efficient design of the non-access stratum (NAS) could also help reduce E2E latency. For example, integrating NAS and access stratum (AS) could reduce the control signaling required to set up and maintain a data connection, which can reduce the E2E latency. Alternatively, developing NAS protocols better tailored to new use cases could also yield a similar result.

Massive Number of Connections

The number of connected devices is expected to increase between 10- and 100-fold beyond 2020 [3]. These will range from devices with limited resources that require only intermittent connectivity for reporting (e.g., sensors) to devices that require always-on connectivity for monitoring and/or tracking (e.g., security cameras, transport fleet). In addition to the sheer number of connected devices, a challenge is to support the diversity of devices and service requirements in a scalable and efficient manner.

A combination of advances in air interface design, signaling optimization, and intelligent clustering and relaying techniques can all contribute to support hyperconnectivity. For instance, using one device as a gateway or relay to aggregate traffic from multiple devices can reduce the signaling load on the network. More efficient protocols that combine AS and NAS also reduce the signaling burden. Moreover, contention-based and connectionless access procedures can be used to efficiently support MTC applications that only require intermittent connectivity to transmit small packets.

Not all devices may be equipped with high-precision devices to cope, for example, with tight synchronization to maintain orthogonality of signals in a multiple access environment when new numerology is introduced to reduce latency. To mitigate this, new waveforms such as FBMC, which can suppress out-of-band emission to reduce interference under an asynchronous environment, can be explored [12]. FBMC also has a potential to cope better than OFDM with doubly dispersive channels when both the transmitting and receiving endpoints are moving (e.g., in a V2V application).

In addition, supporting devices with limited resources such as sensors will require advances in battery and energy harvesting technologies on one hand and efficient signaling and data transmission protocols on the other. For instance, robust medium access techniques combining both control and data transmission could be explored.

Cost

Connectivity is seen as an important enabler for socio-economic development. Therefore, it is important to reduce the infrastructure cost as well as the costs associated with their deployment, maintenance, management, and operation to make connectivity a universally available, affordable, and sustainable utility. The challenge for the design of 5G is that huge improvements are needed to address the new requirements, but customers are not willing to pay proportionally. In effect, 5G should be a network (RAN, core, backbone routers, and backhaul) that addresses all the new requirements at a cost that will make service provisioning sustainable.

Solving the capacity and data rate challenges with network densification could be very expensive in terms of equipment, maintenance, and operations. One way to reduce equipment cost is to minimize the number of functionalities at the base station. This could be done by implementing only layer 1/2 (L1/L2) functionalities in the base station and moving higher-layer functionalities to a network cloud that serves many base stations. Reducing the number of functionalities results in simpler base stations, which could be deployed by users and remotely or autonomously managed to reduce deployment and operation costs.

Energy consumption is a significant operations cost driver, with the RAN estimated to consume 70–80 percent of the energy requirements [13]. Therefore, intelligent energy management techniques, especially in the RAN, could provide a viable means to reduce overall network operations costs. Energy-efficient hardware design, low-power backhaul, and intelligent energy management techniques, especially in ultra-dense networks, to put base stations to sleep when not in use can all contribute to reducing the cost of operating a 5G network [13].

NFV and SDN are also viable enablers to reduce costs. NFV decouples network functionality from dedicated hardware and promotes implementation of functionality in software on general-purpose IT hardware operated according to a cloud model [14]. SDN decouples C- and U-planes of network devices, and provides a logically centralized network view and control, which facilitates transport network optimization. These technologies will make the network more flexible as new functionality can be introduced with simple software upgrades, and more sophisticated algorithms can be employed to manage the network from a holistic viewpoint. Moreover, pooled hardware resources can be shared among multiple functions, thus realizing multiplexing gains and lowering the amount of necessary hardware. The flexibilities enabled by NFV and SDN can make the network quick to deploy and more adaptable, and reduce time to market for new services.

QoE

Quality of experience describes the subjective perception of the user as to how well an application or service is working. Quality of experience is highly application- and user-specific, and cannot be generalized. For example, the QoE of video applications depends on the quality of the encoded and delivered video in the context of the display on which the video is shown. Delivering an application with too low QoE leads to user dissatisfaction, whereas too high QoE unnecessarily drains resources on both the user

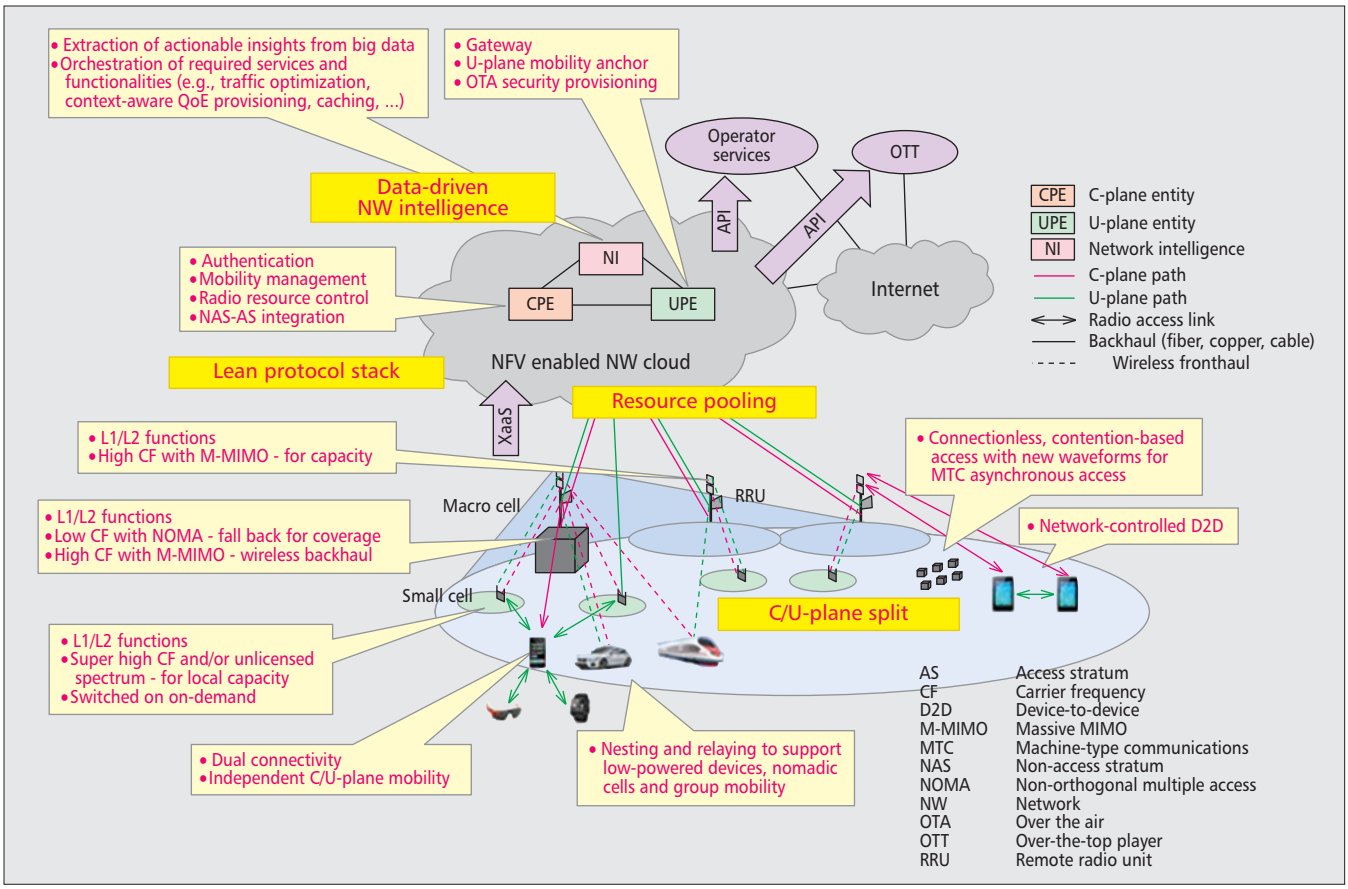


Figure 2. 5G mobile network vision and potential technology enablers.

(e.g., device battery) and operator (e.g., radio and transport network resource, base station power) sides. Hence, a challenge for 5G is to support applications and services with an optimal and consistent level of QoE anywhere and anytime.

Despite the diversity of QoE requirements, providing low latency and high bandwidth generally improves QoE. As such, most enablers mentioned previously can improve QoE. Additionally, traffic optimization techniques can be used to meet increasing QoE expectations. Furthermore, installing caches and computing resources at the edge of the network allows an operator to place content and services close to the end user. This can enable very low latency and high QoE for delay-critical interactive services such as video editing and augmented reality.

Better models that describe the relationship of QoE to measurable network service parameters (e.g., bandwidth, delay) and context parameters (e.g., device, user, and environment) are also emerging. Big data, including information from sensors (e.g., on the device) and statistical user data, can be used intelligently with such models to more precisely assess the QoE expected by a user and determine the optimal resources to use to meet the expected QoE. SDN can then be used to flexibly provision the necessary resources.

Besides the mobile network, advances in the fixed network and potential convergence of the fixed and mobile networks are also needed to address the challenges highlighted above. However, specific discussions related to the fixed network and convergence of the mobile and fixed networks are outside the scope of this article.

5G Mobile Network Architecture Vision

Figure 2 illustrates a 5G mobile network architecture that utilizes the enablers discussed previously. The key elements in the architecture are summarized below:

- Two logical network layers, a radio network (RN) that provides only a minimum set of L1/L2 functionalities and a network cloud that provides all higher layer functionalities
- Dynamic deployment and scaling of functions in the network cloud through SDN and NFV
- A lean protocol stack achieved through elimination of redundant functionalities and integration of AS and NAS
- Separate provisioning of coverage and capacity in the RN by use of C/U-plane split architecture and different frequency bands for coverage and capacity
- Relaying and nesting (connecting devices with limited resources non-transparently to the network through one or more devices that have more resources) to support multiple devices, group mobility, and nomadic hotspots

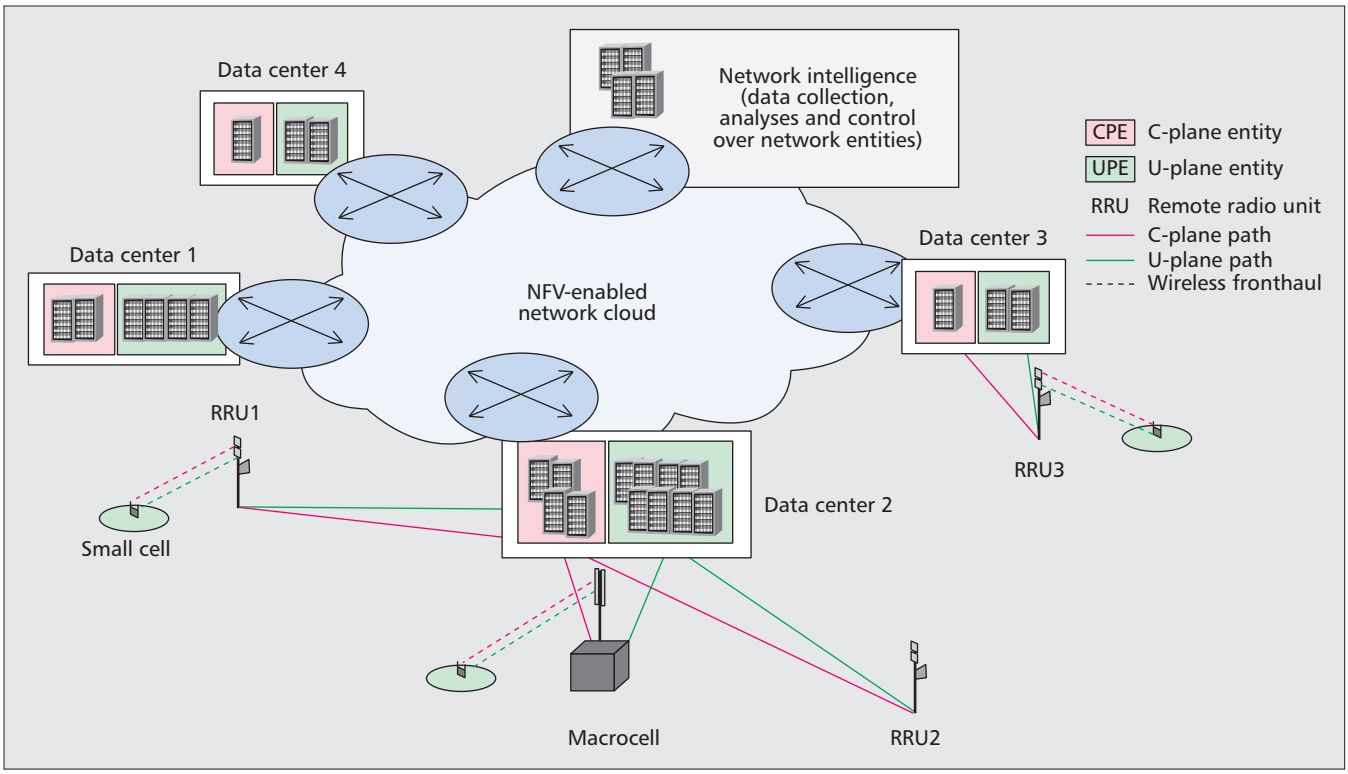


Figure 3. Realization of a 5G network cloud. The network cloud is a logical entity with physical realization that can be tailored to meet specific needs.

- Connectionless and contention-based access with new waveforms for asynchronous access of massive numbers of MTC devices
- Data-driven network intelligence to optimize network resource usage and planning

Logical Network Layers: Radio Network and Network Cloud

The network architecture consists of only two logical layers: a radio network and a network cloud. Different types of base stations and RRUs performing a minimum set of L1/L2 functions constitute the radio network. The network cloud consists of a U-plane entity (UPE) and a C-plane entity (CPE) that perform higher-layer functionalities related to the U- and C-plane, respectively (Fig. 2).

As shown in Fig. 3, the physical realization of the network cloud could be tailored to meet various performance targets. For example, instances of UPEs and CPEs could be located close to base stations and RRUs to meet the needs of latency-critical services. To support latency-critical services, for example, it may be better to connect RRU3 to a small nearby data center (data center 3) rather than a large data center farther away (data center 2). On the other hand, RRU1 may be connected to a large data center located farther away (data center 2) rather than a nearby small data center (data center 1) if support for latency-critical services is not required. Such flexibility allows the operator to deploy both large and small data centers to support specific service needs.

Such architecture simplifies the network and facilitates quick, flexible deployment and management. Base stations would become simpler and consume less energy due to the reduced functionalities, thereby making dense deployments affordable to deploy and operate [15, 16]. Additionally, the network cloud allows for resource pooling, reducing overprovisioning and underutilization of network resources.

Dynamic Deployment and Scaling of Network Functions with SDN and NFV

By employing SDN and NFV, CPE and UPE functions in the network cloud can be deployed quickly, orchestrated and scaled on demand. For instance, when a local data center is unable to cope with a flash crowd (e.g., due to a local disaster), additional capacity can be borrowed quickly from other data centers. In addition, resources within a data center can be quickly shifted to support popular applications simply by adding additional instances of the required software.

Besides this application-level flexibility, the use of a cloud infrastructure also provides flexibility with respect to the available raw processing capacity. Spare cloud resources can be lent out when demand is low, whereas additional resources can be rented through infrastructure as a service (IaaS) business models during peak hours. Furthermore, a broad range of "as a service" business models based on providing specific network functionalities as a service (i.e., XaaS) could also be envisioned. The complete or specific parts of the network could be provided to customers (e.g., network operators, OTT players, enterprises) that have specific requirements, for example in a "mobile network as a service" or "radio network as a service" model. "UPE/CPE/NI as a service" models, where spe-

cific core network functionalities (Fig. 2) of the mobile network are provided a la carte as a service, could also be envisioned. Last but not least, parts of the platform could be rented out to third parties like OTT players to enable the provision of services and applications that require extremely low latency to end users. Besides the XaaS business models that could be facilitated, the flexibility of a cloud, coupled with SDN and NFV technologies, also makes the network easier, faster, and cheaper to deploy and manage.

Lean Protocol Stack

With virtualization, interfaces between network functionalities become interfaces between software. Two separate protocols for the C-plane may no longer be relevant if both NAS and AS protocols can be virtualized. Under a unified cloud paradigm, the NAS and AS protocols can be integrated into a single protocol, removing redundant functionality. In current LTE, for example, the NAS ServiceRequest and RRC ConnectionRequest messages are concatenated, but these could be merged into a single message in a future cloud-based and virtualized network. Similarly, some procedures related to mobility management, session management, and security can potentially be removed. As an example, the connection establishment procedure can be significantly simplified by requiring a handshake only between the peer entities of a single protocol. This in turn will realize faster connection establishment. Bearer-based QoS management could also be replaced by simple IP marking, with proper mechanisms in place to prevent all packets being marked with the highest QoS class.

Similarly for the U-plane, merging of functionalities in the RAN L2 and gateway functionalities in the current core network (CN) can be considered. Virtualization of the U-plane is generally considered to be more difficult than that of the C-plane due to the sheer volume of data to be processed. Virtualization of the RAN L2 protocols can demand significant processing power, as L2 protocols support various features that are dynamic in nature, like dynamic transport block size (according to resource allocation and instantaneous radio condition), segmentation and concatenation of packets, and hybrid automatic repeat request (ARQ). The radio scheduler functionality and advanced features like mMIMO require accurate channel state information (CSI) to be effective. Hence, if such features are to be virtualized, CSI also needs to be delivered to the virtualized entity, potentially imposing significant transport overhead. However, with sufficient advancements in technology and careful selection of functionalities, some of the services provided by L2 can be feasible for virtualization around 2020. In principle, this allows the functionalities provided by different RAN and CN protocols to be merged and a single U-plane entity to provide radio transport services and gateway functionalities. Nevertheless, careful study is needed to determine for which layers such integration can occur.

One feature that can be potentially removed from the U-plane stack is ciphering, since this is increasingly implemented by transport layer security (TLS) over IP. Generally, E2E solutions are more efficient than encrypting segments along the path. However, E2E encryption implies no traffic visibility along the path and makes traffic control in networks difficult. In many operator networks today, intelligent mechanisms such as deep packet inspection (DPI) and caching are used to optimize resource usage and improve QoE. End-to-end encryption would make these intelligent mechanisms dysfunctional. As security of signals transmitted over the air is essential, due to the broadcast nature of radio signals, where to terminate ciphering in the network is an important issue.

Independent Provisioning of Coverage and Capacity with C/U-Plane Split Architecture

Coverage and capacity are provided independently in the RN with a C/U-plane split architecture. Macro and metro base stations provide coverage using licensed spectrum in lower frequency bands and existing cell sites, integrating, for example, NOMA and SIC to boost capacity [8].

Small cell base stations (e.g., Phantom cells [9]) and RRUs provide localized capacity using a combination of licensed and unlicensed spectrum in low and high frequency bands. These cells are deployed indoors and at outdoor hotspots. Advanced schemes (e.g., mMIMO) are also implemented in some RRUs and small cells to boost capacity. Because of the highly variable user and traffic distribution in small cells, they can be put to sleep or switched off completely when they are not needed to save energy. Dynamically switching small cells on and off can provide significant energy savings without degrading network performance [16].

Separating coverage from capacity enables independent mobility of the C-plane and U-plane in areas with overlapping coverage of macro and small cell base stations. In effect, the C- and U-planes for a terminal can take different paths. This requires the terminal to support connectivity to multiple base stations at the same time.

Relaying and Nesting to Support Multiple Devices, Group Mobility, and Nomadic Hotspots

Relays are used as a means to support group mobility (e.g., terminals in a moving vehicle) and nomadic hotspots. In such scenarios, all transmissions within the group are aggregated at one or more entities (e.g., a small cell) and relayed to the network through a wireless backhaul that connects to the network cloud (Fig. 2). Devices with limited resources, such as low-powered wearable devices, connect non-transparently to the network through one or more devices that have more resources (nesting, Fig. 2). By connecting non-transparently, network paging procedures can be used to initiate connections to such devices, thus reducing signaling traffic and power consumption. Together, relaying and nesting provide support for a huge number of devices with diverse capabilities in a scalable and efficient manner.

Because of the highly variable user and traffic distribution in small cells, they can be put to sleep or switched off completely when they are not needed to save energy. Dynamically switching small cells on and off can provide significant energy savings without degrading network performance.

By connecting non-transparently, network paging procedures can be used to initiate connections to such devices, thus reducing signaling traffic and power consumption. Together, relaying and nesting provides support for a huge number of devices with diverse capabilities in a scalable and efficient manner.

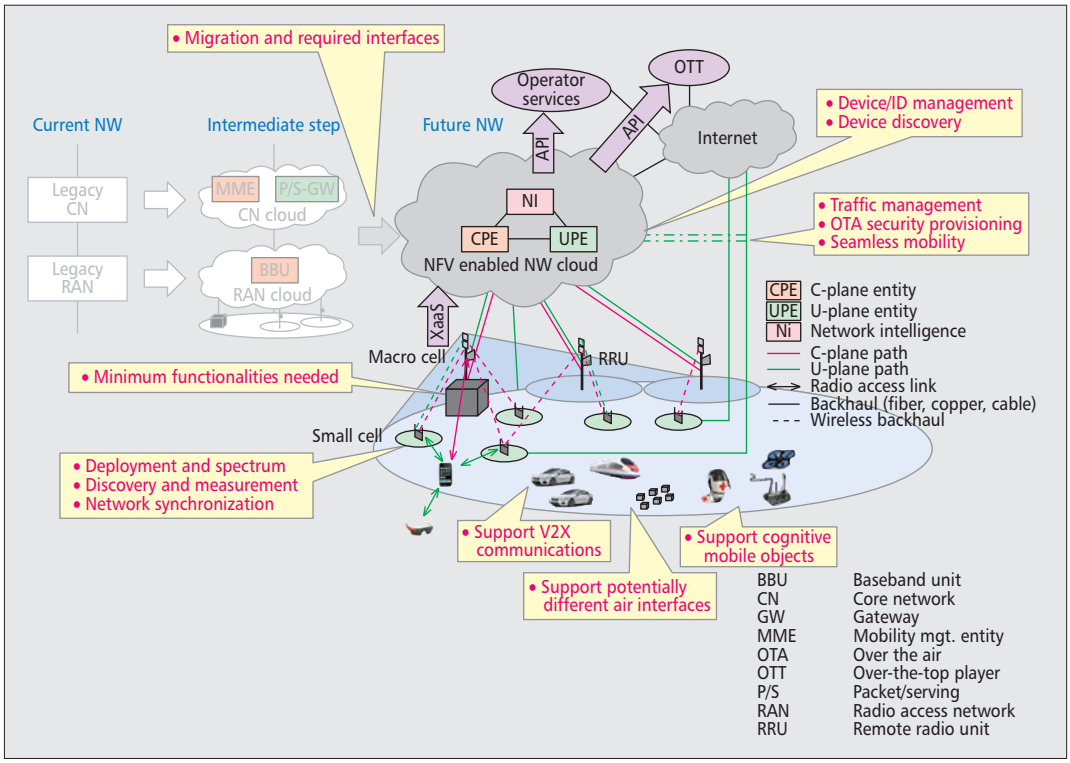


Figure 4. Overview of issues that must be addressed to realize the 5G architecture vision.

Data-Driven Network Intelligence

The architecture allows the network cloud to collect various types of user-centric, network-centric, and context-centric data. The network cloud uses intelligent algorithms to provide real-time insights for efficient resource management, mobility management, local offload decisions (e.g., network-controlled D2D communications), QoE management, traffic routing, and context-aware service provisioning (e.g., geocasting). Furthermore, the aggregated data can provide useful input for network planning. By providing application programming interfaces (APIs) to the network cloud, the collected data can be used in various forms for useful public (e.g., urban planning) and commercial purposes. For example, the APIs can be used to facilitate new businesses based on selling knowledge about network conditions as a service to OTT players, which can allow them to provide consistent service quality to end users.

Issues

Several issues need to be addressed in order to realize the proposed network architecture in particular, and 5G networks in general. Some of these issues are summarized in Fig. 4 and briefly discussed below.

One issue that must be addressed is how legacy networks will interface and interoperate with the new network architecture. One could imagine a migration step where the legacy CN and RAN are migrated to separate cloud platforms during the development phase of 5G (Fig. 4). In order to avoid building parallel networks, it will be essential to specify interfaces and protocols between entities in the legacy clouds and the new network cloud to ensure interoperability.

Another issue is to determine the optimal physical realization of the network cloud to meet performance and cost targets. Whereas centralization of resources could result in savings from pooling, it could also lead to performance bottlenecks, higher latency, and single points of failure. Additional robustness measures will also be needed to avoid devastating impact on service availability if the central entity fails. Moreover, centralization could lead to the need for larger processing and transport capacity at the central entity to process and transport the aggregated traffic, which could diminish the cost savings achieved by pooling. On the other hand, distributing resources could lead to performance improvements and reduced latency, but may be costly due to reduced pooling gains and an increased number of data center locations at corresponding higher operational expenses. Finding the right balance is an important issue.

Ultra-dense small cell deployments will be especially useful for indoor and hotspot environments. As shown in Fig. 5, different deployment options have different implications for the network. In addition to spectrum, backhaul is also an important issue, especially for user deployment. Local breakout may be required for more efficient routing through the user-provisioned backhaul. However, this has implications on the functionalities needed at the small cell base station. For instance, U-plane processing functionalities are needed to support local breakout. Additionally, support for local breakout makes traffic invisible to the network, which affects intelligent QoE provisioning.

Besides the issues highlighted above, seamless mobility provisioning among different types of deployed local and wide-area technologies

	Operator-deployed	User-deployed
Licensed spectrum	Pros • Cell sites fully controlled by the operator • Easier to provide QoE • Advanced resource allocation (RA) techniques become easier to realize Cons • Cost (equipment, deployment, operation) • Limited spectrum • Spectrum license fees Issues • Backhaul provisioning	Pros • Reduced cost (equip., deployment, operation) Cons • Additional operation costs to provide after-service customer support Issues • Regulatory issues • Access control (public or private) • Ensuring QoE, e.g., new mechanisms to control interference (e.g., low Tx power) • Impact of diverse backhaul types on advanced RA techniques (e.g., CoMP) • Provisioning of over-the-air security
Unlicensed spectrum	Pros • Cell sites fully controlled by the operator • Additional spectrum for operators to exploit Cons • Cost (equipment, deployment, operation) • Lack of QoE guarantees Issues • Mechanisms to ensure fair-play (definition and implementation of incentive-compatible spectrum etiquette) • Coexistence with Wi-Fi, Bluetooth, etc. • Backhaul provisioning	Pros • Reduced cost (equip., deployment, operation) Cons • Lack of QoE guarantees Issues • Access control • Mechanisms to ensure fair-play (definition and implementation of incentive-compatible spectrum etiquette) • Coexistence with Wi-Fi, Bluetooth, etc. • Impact of diverse backhaul types on advanced RA techniques (e.g., CoMP) • Provisioning of over-the-air security

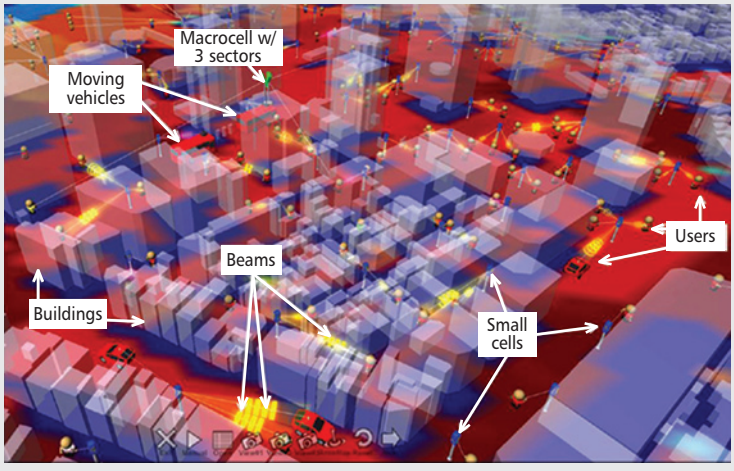
Figure 5. Small cell deployment options and issues.

with potentially different functionalities also has to be addressed to improve the overall QoE for end users. Mechanisms to support simultaneous sessions and seamless session mobility across different access networks will also be required to support consistent QoE for end users. Furthermore, different types of edge networks will also need to be integrated within the 5G network architecture. For instance, the communication needs of cognitive mobile objects (robots2X, drones2X, etc.) will all need to be efficiently supported and integrated in 5G networks. Finally, new paradigms of identity management and charging will need to be developed for 5G, in particular, to cope with the huge number of devices expected to be connected to the network, the diverse use cases, and different edge network topologies.

Initial Proof of Concept

A real-time simulator is used to evaluate the system-level gains when some of the candidate 5G technologies described in the previous section are introduced for downlink transmission. Specifically, the gains from the hybrid usage of macrocells at lower frequency bands and small cells at higher frequency bands, together with mMIMO, are demonstrated.

Figure 6 shows the deployment environment studied, which consists of buildings, moving vehicles, users, macro base stations, and a dense deployment of small cell base stations. A seven-cell model is assumed with an inter-site distance of 500 m. Each macrocell has three sectors, and each sector has 30 outdoor users (i.e., penetration loss = 0 dB). A 3 km/h user speed is assumed. Ray tracing is applied using the vertical plane launch (VPL) method to emulate a real propagation environment of a 750 m × 750 m dense urban area in Shinjuku, Tokyo. The baseline system



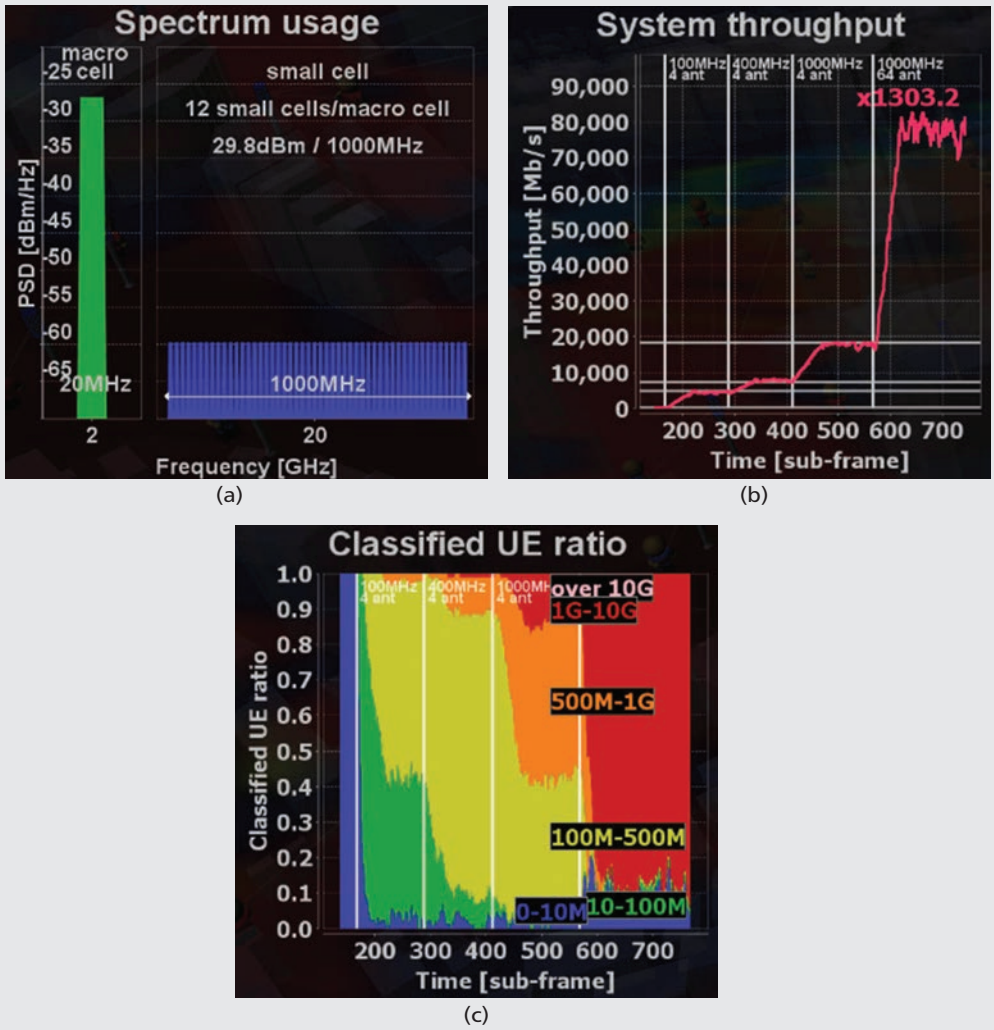
Color code of users and the ground indicates achievable user data rate:
Blue: 0–10 Mb/s; Green: 10–100 Mb/s; Yellow: 100–500 Mb/s;
Orange: 500 Mb/s–1 Gb/s; Red: 1–10 Gb/s; Pink: over 10 Gb/s

Figure 6. The deployment environment of the 5G real-time simulator.

consists of LTE-based macrocells using 20 MHz bandwidth at 2 GHz. Each macrocell uses two transmit (Tx) antennas. An antenna gain of 14 dBi and a total transmit power of 49 dBm are assumed for each macrocell base station. For evaluating the gains of network densification and wideband transmission at higher frequency bands, 12 small cells are deployed per sector. Each small cell uses 1 GHz bandwidth at 20 GHz. The number of Tx antennas per small cell is 64. An antenna gain of 5 dBi and a total transmit power of 30 dBm are assumed for each small cell base station. The number of receive antennas at the user terminal is 4 at both 2 GHz and 20 GHz.

For 2 × 4 MIMO transmission in macrocells, single-user MIMO is applied based on implicit

New paradigms of identity management and charging will need to be developed for 5G, in particular, to cope with the huge number of devices expected to be connected to the network, the diverse use cases and different edge network topologies.



Color code for Figure 7c indicates the fraction of users able to achieve a particular range of data rate: Blue: 0–10 Mb/s; Green: 10–100 Mb/s; Yellow: 100–500 Mb/s; Orange: 500 Mb/s–1 Gb/s; Red: 1–10 Gb/s; Pink: over 10 Gb/s

Figure 7. Spectrum usage and performance evaluation results of the 5G real-time simulator: a) power spectrum density (dBm/Hz) vs. frequency (GHz); b) throughput (Mb/s) vs. time (subframe); c) classified UE ratio vs. time (subframe).

CSI feedback using the LTE Release 8 codebook. For 64 × 4 mMIMO transmission in small cells, the CSI of users is assumed to be perfectly known at the small cell base station side, and Hermitian precoding is applied for multi-layer transmission. In order to improve both cell coverage (by beamforming gain) and spectrum efficiency (by spatial multiplexing gain) of small cells, single-user MIMO and multi-user MIMO dynamic switching (up to 4 users) and rank adaptation (up to 4 layers/user) are introduced. Proportional fairness scheduling is applied to allocate frequency/time resources to users at macrocells and small cells disjointly. Note that no intercell interference coordination (ICIC) is applied among either macrocells or small cells.

The performance of the candidate technologies are shown in Fig. 7. Figure 7a illustrates the spectrum usage for macrocells and small cells. It can be seen that the power spectrum density (PSD) becomes lower as the spectrum bandwidth is extended to 1 GHz for small cells. In Figs. 7b and 7c, the x-axis (time [subframe]) refers to the number of subframes being processed and also the time in milliseconds (one subframe = 1 ms). The system throughput per subframe of a 500 m × 500 m area is shown in Fig. 7b, which demonstrates that compared to a macro-only 3GPP Release 8 LTE deployment, around 1300× system throughput gains are achieved by a combination of dense deployment of small cells, using large bandwidths at higher frequency bands and employing mMIMO techniques at small cells. By simulating each of the candidate 5G technologies above, we see the 1300× system throughput gains as the combination of almost 50× from bandwidth extension from 20 MHz to 1 GHz, 4× from antenna densification by adding 12 small cells per sector, and around 6.5× from mMIMO by introducing 64 × 4 mMIMO with single-user MIMO and multi-user MIMO dynamic switching.

Finally, Fig. 7c shows the classified UE ratio, which gives the fraction of users who are able to achieve a particular range of data rate. It can be seen from this that more than 90 percent of users are able to achieve data rates in excess of 1 Gb/s (i.e., the red color zone expanded to below 0.1) with such a network. These initial results demonstrate the potential of network densification using small cells, bandwidth extension in higher frequency bands, and mMIMO at small cells to address the capacity and data rate challenges of 5G networks.

Conclusions

The important challenges that must be addressed by 5G networks have been highlighted: higher capacity, higher data rate, lower E2E latency, massive device connectivity, reduced capital and operation cost, and consistent QoE provisioning. A 5G architecture vision to address some of those challenges is presented and a two-layer architecture proposed, consisting of a radio network and a network cloud. The proposed architecture integrates various enablers such as small cells, massive MIMO, C/U-plane split, NFV, and SDN. The main concepts can be summarized as follows:

- Ultra-dense small cell deployments on licensed and unlicensed spectrum, under C/U-plane split architecture, to address capacity and data rate challenges
- NFV and SDN to provide flexible network deployment and operation, with integrated AS and NAS features
- Intelligent use of network data to facilitate optimal use of network resources for QoE provisioning and planning

Initial proof of concept investigations suggest more than 1000 times throughput gains compared to a macro-only 3GPP Release 8 LTE deployment are achievable by a combination of dense deployment of small cells, using large bandwidths at higher frequency bands and employing massive MIMO techniques at small cells. Nevertheless, some of the components highlighted in the system concept have mutual conflicts when details are considered. Hence, how to balance the pros and cons of each aspect needs to be carefully studied. Further investigations are necessary, particularly in the following areas: suitable techniques for use in small cells in different frequency regimes; how to incorporate small cells with NFV and SDN in a cost-effective manner; and intelligent algorithms that better utilize the available network resources to provide a consistent end-user QoE.

References

[1] Qualcomm, "The 1000x Mobile Data Challenge," White Paper, Nov 2013.
[2] NSN, "Signaling is Growing 50% Faster than Data Traffic," White Paper, 2012.
[3] METIS, "Scenarios, Requirements and KPIs for 5G Mobile and Wireless System (Deliverable D1.1)," May 2013.
[4] "Advanced 5G Network Infrastructure for the Future Internet — Public Private Partnership in Horizon 2020," 2013.
[5] 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.
[6] E. G. Larsson *et al.*, "Massive MIMO for Next Generation Wireless Systems," May 2013.
[7] S. Gollakota, S. Perli, and D. Katabi, "Interference Alignment and Cancellation," *ACM SIGCOMM Comp. Commun. Rev.*, 2009.
[8] A. Benjebbour *et al.*, "System-Level Performance of Downlink NOMA for Future LTE Enhancements," *IEEE GLOBECOM*, 2013.
[9] H. Ishii, Y. Kishiyama, and H. Takahashi, "A Novel Architecture for LTE-B: C-plane/U-Plane Split and Phantom Cell Concept," *IEEE GLOBECOM Wksps.*, 2012.
[10] G. Fodor *et al.*, "Design Aspects of Network Assisted Device-to-Device Communications," *IEEE Commun. Mag.*, vol. 50, no. 3, 2012, pp. 170–77.
[11] G. P. Fettweis, "A 5G Wireless Communications Vision," *Microwave J.*, Dec 2012.
[12] B. Farhang-Boroujeny, "OFDM Versus Filter Bank Multicarrier," *IEEE Signal Proc. Mag.*, vol. 28, no. 3, May 2011, pp. 92–112.
[13] EARTH Project Work Package 2, "Deliverable D2.1: Economic and Ecological Impact of ICT," https://www.ict-earth.eu/publications/deliverables/deliverables.html, 2011.
[14] "Network Functions Virtualisation - Introductory White Paper," SDN and OpenFlow World Congress, Darmstadt, Germany, Oct 2012.
[15] EARTH Project Work Package 4, "Deliverable D4.3: Final Report on Green Radio Technologies," https://www.ict-earth.eu/publications/deliverables/deliverables.html, 2012.
[16] E. Ternon *et al.*, "Database-Aided Energy Savings in Next Generation Dual Connectivity Heterogeneous Networks," *IEEE WCNC*, Istanbul, Turkey, Apr 2014.

Biographies

Patrick Agyapong is a researcher with the wireless research group at DOCOMO Communications Laboratories Europe. His research focuses on designing incentive-compatible algorithms, protocols, and architectures to support next generation mobile communication needs, spanning the areas of resource management, content distribution, network architecture, and business strategy. He holds a Ph.D. in engineering and public policy from Carnegie Mellon University, Pittsburgh, Pennsylvania. He also holds an M.Sc. in electrical engineering and a B.Sc. in electrical engineering and computer science, both from Jacobs University Bremen, Germany.

Mikio Iwamura is a director of the wireless research group at DOCOMO Communications Laboratories Europe. He received his Ph.D. and M.Sc. degrees from King's College London in 2006 and the Science University of Tokyo in 1998, respectively. Before his current role, he was deeply engaged in LTE standardization, with over 300 contributions to 3GPP. He has over 100 patents internationally, and has published over 20 technical journals and conference papers.

Dirk Staehle is a manager at DOCOMO Communications Laboratories Europe. He is responsible for the standardization team contributing to the 3GPP SA2 and ETSI NFV standardization groups. He received his Ph.D. from the University of Würzburg in 2004 and continued as an assistant professor, leading the wireless network research group before joining DOCOMO in 2011. His research activities include NFV, machine-type communication, application and QoE-aware traffic and resource management, and radio network planning.

Wolfgang Kiess studied at the Universities of Mannheim and Nice, and holds a diploma in business information systems from the University of Mannheim and a Ph.D. in computer science from the University of Düsseldorf. He is the leader of the virtualization research team at DOCOMO Communications Laboratories Europe, focusing on cellular core network virtualization, cloud computing, and 5G.

Anass Benjebbour [SM] obtained his Ph.D. and M.Sc. in telecommunications in 2004 and 2001, respectively, and his Diploma in electrical engineering in 1999, all from Kyoto University, Japan. He is currently an assistant manager of the 5G team within NTT DOCOMO, Inc. He served as 3GPP RAN1 standardization delegate during LTE Release 11, as secretary of the IEICE RCS conference, and Editor for *IEICE Communications Magazine*. He is a Senior Member of IEICE.

Initial proof of concept investigations suggest more than 1000 times throughput gains compared to a macro-only 3GPP Release 8 LTE deployment are achievable by a combination of dense deployment of small cells, using large bandwidths at higher frequency bands and employing massive MIMO techniques at small cells.

A New Control Plane for 5G Network Architecture with a Case Study on Unified Handoff, Mobility, and Routing Management

Volkan Yazıcı, Ulaş C. Kozat, and M. Oğuz Sunay

ABSTRACT

The tremendous growth in wireless Internet use is showing no signs of slowing down. Existing cellular networks are starting to be insufficient in meeting this demand, in part due to their inflexible and expensive equipment as well as complex and non-agile control plane. Software-defined networking is emerging as a natural solution for next generation cellular networks as it enables further network function virtualization opportunities and network programmability. In this article, we advocate an all-SDN network architecture with hierarchical network control capabilities to allow for different grades of performance and complexity in offering core network services and provide service differentiation for 5G systems. As a showcase of this architecture, we introduce a unified approach to mobility, handoff, and routing management and offer connectivity management as a service (CMaaS). CMaaS is offered to application developers and over-the-top service providers to provide a range of options in protecting their flows against subscriber mobility at different price levels.

INTRODUCTION

Mobile operators are faced with several major challenges such as the unprecedented increase in wireless traffic, losing existing and new revenue sources to over-the-top (OTT) providers, and increasing capital as well as operational expenditures to serve the demand. To meet these challenges, many mobile service providers (MSPs) are heavily pursuing cloudification opportunities, mainly in the form of network function virtualization (NFV) [1]. NFV aims to move some or all of the functions of the mobile network from dedicated hardware platforms to virtual machines running on generic hardware. NFV promises reduced expenditures, agility and flexibility, and the capability for MSPs to launch new network services to seize new market opportunities in timescales that OTT providers can achieve. Virtualization and pooling of baseband processing in the base stations is one example of NFV, and is commonly referred to as the cloud radio access network (C-RAN) in the literature [2].

Moving toward cloudification, a brute force approach would be to keep the current network architecture and simply run core network nodes (xGSN, S-GW, P-GW, MME, PCRF, HSS, etc.) as well as service platforms (e.g., IMS) in a virtualized data center environment (referred to as a telco cloud) [3]. We argue that while this is a step in the right direction toward easing the problems of MSPs, it is not sufficient. Ultimately, a mobile network is concerned with the forwarding of flows to and from mobile user equipment (UE) via a chain of network functions. While NFV brings programmability, agility, and flexibility to the realization of individual functions, it is essential that a programmable, agile, and flexible realization of individualized flow control that orchestrates flows across different chains of functions is also present in 5G systems so that MSPs can quickly create and deploy new revenue-generating services.

We advocate that a key differentiator of 5G systems from 4G will be in how we architect and orchestrate the overall system control to realize the benefits of cloudification while taking the full advantage of the transport capacity distributed over a large geographical area in the form of base stations, switches, routers, fiber links, microwave links, etc. We envision fully decoupled, independently scalable and programmable user and control planes for 5G. In this respect, software defined networking (SDN) is a natural architectural choice for 5G systems. In SDN architectures, complex control plane functions (CPFs) are removed from forwarding elements and placed *behind* a logically centralized controller. Thus, the SDN approach simplifies the forwarding elements and ships complex CPFs to physical or virtual servers running in a data center. The controller collects the distributed net-

Volkan Yazıcı and M. Oğuz Sunay are with Özyeğin University.

Ulaş C. Kozat is with Docomo Innovations, Inc.

0163-6804/14/$25.00 © 2014 IEEE

work state on behalf of CPFs and provides them direct centralized access to raw network state or their abstracted forms. Based on this centralized view, CPFs dictate decisions on how packets are processed, pipelined, and forwarded on one or more data plane nodes (DPNs) by sending instructions back to the controller, which in return sanitizes and translates these instructions for individual forwarding elements using an open standard (e.g., OpenFlow, NetConf) or proprietary interface supported at individual DPNs.

Mobile networks are composed of two components: the RAN and the core network (CN). While the RAN provides connectivity of the UEs to the network via base stations (eNBs), the CN provides paths between eNBs and various services as well as outside networks. Then, considering the mobility of UEs, the delay constraints associated with various control functionalities of the mobile network are significantly different. In this article, while we advocate an all-SDN network architecture for 5G, we ascertain that an interworking set of hierarchical controllers, as opposed to a single centralized controller, is necessary to handle such variations in the delay constraints. The hierarchy of controllers allow not only for locally optimized control decisions within the network, but also a new dimension in service provisioning, where control of a given service at different hierarchies corresponds to different grades of service.

The rest of this article is organized as follows. First, we give an overview of the control plane architecture in the current cellular system and discuss why rethinking is necessary for 5G. Then we present an overview of the existing SDN-based architecture proposals for the CN and RAN from the literature. Next, we introduce a new network architecture that has the following characteristics:

- A hierarchy of controllers are deployed to provide flow-based service grades for different CPFs.
- Overlay routing via tunnels is completely eliminated.
- Handoff decisions, mobility management, and end-to-end routing are unified under a single CPF.
- New instructions for southbound interface are specified to orchestrate control functions programmatically.
- Multiple CPFs for the same functionality may be described to provide such functions as a cloud service to open new revenue venues for MSPs.

We then introduce a use case for unified mobility and routing management in the new architecture, summarize some of our experimental findings that show how we can trade off wireless connectivity performance with network operational cost on a per flow basis. Finally, we draw some conclusions.

Why the 4G Control Plane Needs to Change

The 4G cellular system has evolved from its third generation counterpart. For this reason, the RAN and CN components are usually referred to as Long Term Evolution (LTE) and System Architecture Evolution (SAE), respectively. Together, they form the Evolved Packet System (EPS) [4]. The EPS is an all-IP network supporting only packet-switched connectivity. All radio related functions are pushed down to the eNBs in the EPS to increase delay performance when reacting to the changes in the wireless environment. EPS has clean separation of the user and data planes in the CN to allow the independent scaling of the two planes.

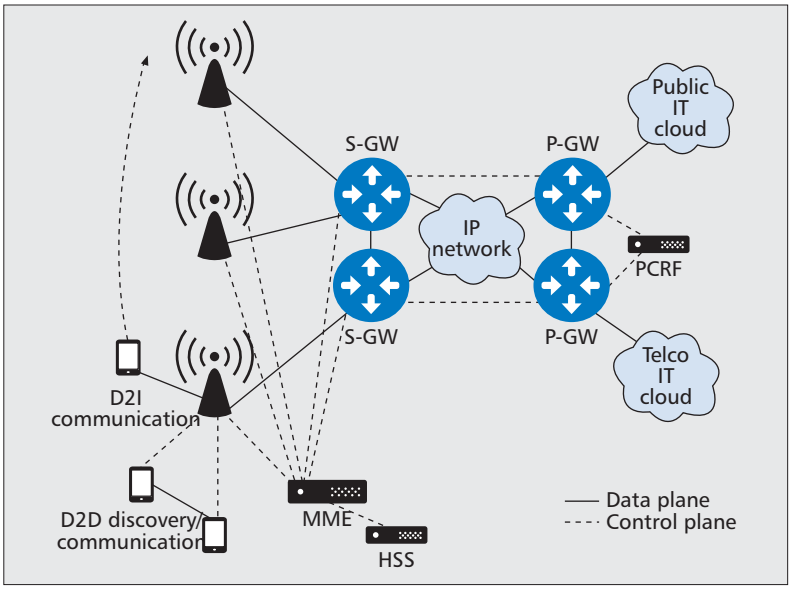


Figure 1. Current EPS network architecture.

The control operations in the EPS are functionally split between the RAN and the CN. The RAN has a single element, the eNB. The eNB is responsible for admission control; inter-cell radio resource management (RRM); radio resource block (RB) control and scheduling; and handoff management. In contrast, the CN consists of several elements: mobility management entity (MME), serving gateway (S-GW), packet gateway (P-GW), home subscriber server (HSS), and policy and charging rules function (PCRF). The MME is mainly responsible for paging and mobility management, but is also involved in bearer management, admission control, subscription management, and so on. The HSS is involved in subscription management. The PCRF, S-GW and P-GW are all involved in bearer establishment, maintenance, and quality of service (QoS). For instance, the PCRF dictates QoS policies and charging for individual flows and subscribers, while enforcement of these policies through mapping flows to bearers, performing packet filtering and metering are tasks of the P-GW. The P-GW is also responsible for IP address assignment for the UEs. The S-GW is involved in buffering packets for idle mode subscribers and triggering paging through the MME. The current EPS network architecture is illustrated in Fig. 1.

The massive growth in traffic volume as well as the volume of connected devices necessitates an evolution of LTE toward 5G. The latest significant step in this evolution is Third Generation (3GPP) Release 12, where the new use case

While LTE has achieved significant gains in lowering the MSP expenditures and increasing end-user experience with higher data rates and lower latencies, a new thinking of the cellular network architecture is necessary toward 5G.

of device-to-device (D2D) communication is being studied for the first time under the title proximity services (ProSe) [5]. The 3GPP defines D2D communication as communication between two nearby UEs directly, without routing through the EPC. D2D communication may or may not involve routing through the local eNB. D2D communication will improve spectrum and energy efficiency of the overall system while increasing the throughput and end-to-end delay performance for D2D links [6]. D2D-enabled LTE is also aimed at mission-critical communication systems that must function when cellular networks are not operational.

In 3GPP, D2D studies are centered around two fundamental types of operations:

- D2D discovery
- D2D communication

While D2D discovery aims to identify other UEs in close proximity for possible D2D communication, D2D communication is the actual direct link for data transfer. For both operations, it is essential to architect an efficient, fast, and robust control plane so that D2D terminals are time synchronized to the network, and are allocated and subsequently scheduled for the necessary resources.

D2D discovery should be designed in such a way that the UEs wake up only to listen for potential D2D partners. This is best achieved by allocating static wireless resources to the discovery operation in the network. Conversely, D2D communication should utilize resources only when necessary so that spectrum efficiency is maintained in the network. This requires dynamic network-controlled allocation of resources to the D2D link. For both D2D discovery and communication, uplink resources are favored [6].

D2D discovery and communication operations should be operational beyond the coverage of the network. In the absence of network guidance, the pool of UEs that are candidates for D2D communications may use either an ad hoc or cluster-head-based control mechanism. If an ad hoc mechanism is employed, each UE controls itself, and discovery and communication may utilize one of the well-known random medium access control (MAC) protocols such as carrier sense multiple access (CSMA) [6]. If, on the other hand, a cluster-head-based control topology is preferred, one UE assumes a master role and performs all of the control operations that local eNBs provide under network coverage. Cluster-head-based control is also suitable for D2D-based range extension within the network, where the selected cluster head acts as a relay to convey network control and communication data to out-of-coverage UEs in its proximity.

While LTE has achieved significant gains in lowering MSP expenditures and increasing end-user experience with higher data rates and lower latencies, fresh thinking on the cellular network architecture is necessary toward 5G. The explosive demand in wireless data is driving a heterogeneous network (HetNet) paradigm with a large number of small cells as well as allowing D2D connectivity and resulting in potentially a significantly more complex RAN and CN control, pushing up the deployment and operational costs substantially. 5G networks will likely experience a continuous deployment of small cells based on changes in local user demand. As it is not possible to re-architect the network every time it gets denser, an agile and configurable solution is needed. Furthermore, it is very likely that the local demand for the data and control planes of 5G networks will grow at different paces, necessitating an independently scalable solution. As user demand increases, so does user intolerance to underperforming applications. 5G networks should be able to dynamically steer or reprioritize individual traffic flows based on network-wide orchestration utilizing big data analytics to ensure user satisfaction. For improved performance, 5G networks will also require a more coordinated approach to RAN technologies, as already exemplified in the latest release LTE systems in the form of coordinated multipoint (CoMP) transmission and intercell interference coordination (ICIC) where multiple eNBs serve a UE in coordination, and coordinate their transmissions to minimize their interference to one another, respectively. Such coordinated technologies will potentially be better realized via a logically centralized control plane spanning over the eNBs of a given geography as opposed to the current distributed approach of LTE.

Furthermore, reduction of the costs of such a network will need to come in two major fronts:

- *Cloudification*: Virtualization of various network functions will enable the realization of a multitude of functionalities in a virtualized data center environment, eliminating the need for specialized hardware.
- *Programmability*: A programmable network will enable MSP-led innovations of new control applications and the corresponding chain of services to provide control differentiation of different flows even for some of the most fundamental network operations such as mobility management.

Then an agile and flexible 5G architecture with perfectly decoupled data and control planes where virtualized network functions as well as data flows can be orchestrated programmatically is necessary. We argue that the SDN framework is the ideal candidate for such an architecture. After providing a brief review of the existing SDN proposals for cellular networks, we introduce the proposed architecture.

Overview of Existing SDN Proposals for CN and RAN

There have been a number of studies in the literature that detail how the current 3GPP data plane may be realized using an SDN framework. For instance, in [3], Kempf *et al.* propose supporting the tunneling protocol between the gateway nodes and eNBs as an extension to the OpenFlow Switch (OFS) specification [7]. Furthermore, they envision that the transport fabrics between the tunnel endpoints are also OFS, centralizing the control over the paths that flow between the P-GW and eNBs. As a result, both the S-GW and P-GW can be split from their data plane functions and instead run purely as control plane applications. However, the proposed architecture does not mention how com-

plex policies can be enforced on the extended OFS, a crucial component for carrier-grade SDN. MobileFlow introduces a complementary SDN architecture with a logically centralized MobileFlow controller for mobility management with legacy equipment support in addition to the OpenFlow controller for routing [8]. While MobileFlow provides a promising architecture, it does not specify the interplay of the two controllers in detail for different network control operations or mention how the control of the RAN might be integrated into the architecture.

The Softcell architecture highlights possible bottleneck issues due to many functions hosted on P-GWs, and advocates a cleaner separation between the data and control plane functions [9]. In the proposal, local control agents interact with the more centralized controller to resolve the timescale issues in control loops. The authors advocate that the current OpenFlow model is not sufficient to perform useful functions such as deep packet inspection (DPI) and header compression on the path.

The OpenRoads architecture discusses the necessity of coordination between the RANs of different radio access technologies (RATs) to enable seamless inter-RAT handoffs [10]. OpenRadio discusses the benefits of a software programmable data plane for the RAN through the decomposition of wireless protocols into separate processing and decision plane components with a simple programmable application programming interface (API) between them [11]. SoftRAN suggests a logically centralized control plane for the RAN in a given geography composed of a number of eNBs, where parts of the control remain at the eNBs [12]. Specifically, operations of handoffs and power control are handled at a centralized network controller, whereas each eNB controls resource allocation. OpenRF proposes a central controller for coordinated interference management of MIMO-based Wi-Fi networks [13]. The OpenRF controller assigns each flow to an access point (AP) and establishes corresponding interference and coherence vectors. These assignments are conveyed to the APs, which in return combine all assigned coherence and interference vectors to produce precoding vectors that enable transmission of desired flows coherently while nulling any interference these flows may cause to other active flows.

Extending on all of the above reported work, we propose an all-SDN architecture for the mobile network with hierarchial controllers. In the next section we discuss this architecture.

A New Programmable 5G Control Plane Architecture

The 5G network will most likely be heterogeneous in its deployment with densely populated small cells, with device-to-infrastructure (D2I) as well as device-to-device (D2D) links, and operational on a number of different carrier frequencies, ranging from today's cellular bands below 5 GHz to millimeter-waves at 60 GHz and beyond. Evolving from LTE, 5G will likely have an extensive set of adaptive physical layer components relying on large numbers of transmit and receive antennas.

We believe a simpler programmable network architecture will be able to support such a vision. This architecture will completely eliminate specialized and thereby expensive components such as the MME, S-GW, P-GW, and PCRF, as well as tunneling protocols used for overlay routing. Furthermore, this architecture will unify the control of the RAN and CN for the network to allow for flexibility in orchestrated network programmability. While we advocate an all-SDN network architecture, we ascertain that an interworking set of hierarchical controllers as opposed to a single centralized controller is necessary to handle the delay constraints associated with various control functionalities of the mobile network. For instance, scheduling of wireless resources is traditionally based on the channel quality feedback received from users. The coherence time of this feedback is dependent on the carrier frequency as well as the user mobility. At 2 GHz, this is equal to 90 ms and 1.1 ms for a user travelling at 3 km/h and 250 km/h, respectively. At 5 GHz, these values reduce to 36 ms and 0.43 ms, respectively. Considering a 5–10 ms one-way delay in traditional backhaul links [12], centralized control of such a scheduling operation away from the individual eNBs may be feasible for a low-mobility user, but certainly not for a high-mobility user. The proposed architecture provides the flexibility to divide the wireless resources within a geographical area with multiple eNBs into a number of virtual slices and perform scheduling within these slices at different controller hierarchies. While scheduling at the eNB controller allows for reactive utilization of wireless resources, it is conducted with a limited local network view. On the other hand, scheduling conducted at the RAN controller may lead to better prioritization of flows, utilizing a larger network view, at the expense of an increase in the reaction time to network dynamics.

In the proposed architecture, which is depicted in Fig. 2, the core control functions of the wireless network such as connectivity, RAT selection, handoff management, mobility management (MM), C-RRM, QoS, policy, and charging are all realized as applications running on one or more of the hierarchical controllers. Furthermore, multiple control applications for the same functionality may be present in the network, realized at the same or different controller hierarchies. The selection of the control application for a given flow may depend on not only the user identity, flow, connection type, but also user mobility, user observed channel quality, network carrier frequency, mobile phone capability, user billing plan, roaming information, OTT identity, and more that jointly make up the mobile network state. It may also be possible to switch from one control application to another during the lifetime of a flow due to a change in the network state.

As illustrated in Fig. 2, the following controllers are defined in the proposed architecture in decreasing hierarchy: network controller, RAN controller, base station (BS) controller, and UE controller. For a mobile network covering a wide area, it may be desirable to realize the logically centralized network and RAN controllers in a distributed fashion across the topology to provide scalability and increased

An agile and flexible 5G architecture with perfectly decoupled data and control planes where virtualized network functions as well as data flows can be orchestrated programmatically is necessary.

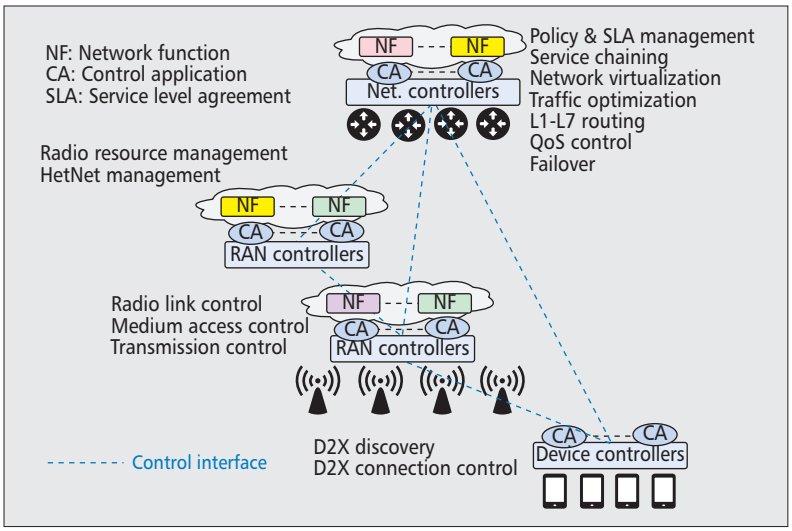


Figure 2. Programmable all-SDN 5G network architecture.

performance. By their nature, the BS and UE controllers are already distributed.

Analogous to the OpenFlow interface [7], the control of the end-to-end network operates over tables of <Match, Action> tuples, sent by the controllers to the forwarders (routers, BSs, and user equipment). However, an extended set of match and action attributes are needed for the mobile network. Any subset of the above described network state may be used as a Match entry. The corresponding Actions may be to select a RAT, schedule or avoid a specific wireless resource, initiate handoff, set the modulation and coding, initiate an automatic repeat request (ARQ) protocol, charge according to a specific policy, initiate CoMP, initiate ICIC, forward on a specific port, pause a flow, resume a flow, limit the data rate and bandwidth, allow/disallow D2D communication, act as a relay, and so forth. An efficient representation of the network state and associated control actions within the <Match, Action> tuple is necessary and requires further investigation.

The control of the network also necessitates close interaction between the hierarchical controllers. In the proposed architecture, a controller at a higher hierarchy may send constraints to a controller at the lower hierarchy using a similar <Match, Action> tuple. Here, the Match entries may include any subset of the network state, and the Action entries include the selection of a control application, disabling or enforcing the joining of a RAT, limiting modulation and coding options, disallowing simultaneous scheduling of the same resource to the matching flows, disallowing handoffs to certain types of BSs, disallowing or only allowing D2D communication, limiting transmit power for a given resource, powering a BS on or off, and so on. Conversely, a controller at a lower hierarchy sends abstracted feedback to a controller at the higher hierarchy upon the higher hierarchy controller's demand or at regular intervals. For example, the RAN controller may disallow the handoff of a high mobility user from a macrocell to a small cell. Similarly, the RAN controller may require the same RBs to be or not to be scheduled for a given user by the BS controllers for CoMP operation. This constraint does not, however, negate the autonomous operation of the scheduler application running on the BS controllers. In return, these controllers may send an abstracted feedback of the users' observed channel quality histories to the RAN Controller, which in turn may use this information to decide when and how to instantiate CoMP or ICIC.

Let us now describe the controllers in the hierarchy. The UE Controller is responsible for the selection of one of many available radio access technologies (RAT) that the device supports subject to the limitations that are imposed locally or by one of the controllers in the higher hierarchy. This way, RAT selection based on subscription, policy and charging on a per-flow basis becomes possible. The UE controller is also responsible for various D2D discovery and communication control operations: push or pull based discovery control, D2D physical layer modulation and coding adaptation, H-ARQ operation, out-of-coverage D2D distributed random access and/or cluster head-based centralized control options for discovery and communication, etc.

One step higher in the control hierarchy is the BS Controller. As described above, we believe that delay-constrained functions such as wireless resource management and scheduling as well as the corresponding adaptive physical layer packet creation need to be controlled close to the UE at the BS Controller for D2I communication. The D2D resource management and synchronization of the in-coverage D2D UEs are also controller by the BS controller. The RAN Controller, which is one step up on the hierarchy, oversees all of the BSs in a given geographical area and thus has the potential to effectively control the C-RRM functionalities. However, the proposed architecture also allows for the C-RRM control at the BS Controller in a distributed fashion, similar to today's LTE solution. It is also possible to invoke one C-RRM control for one flow, and another control for another.

The Network Controller at the top of the hierarchy potentially orchestrates end-to-end QoS provisioning, application-aware route establishment and service chaining, mobility management, policy and charging and it percolates/delegates its decisions on the controllers in the lower levels of the hierarchy.

The hierarchical control of the all-SDN network allows for the realization of a given control operation at different hierarchies, possibly using different control applications and introducing new venues for revenue generation for MSPs. One important example of this flexibility is for connectivity management, which is comprised of mobility management and dynamic route management. This is discussed in the next section.

Case Study: Joint Control of Active Mode Mobility and Route Management

One of the fundamental goals of network control is to ensure that a route for a given flow between two nodes is quickly and effectively

established and maintained. In a wireless network, this problem becomes significantly more complex as one or both of the nodes are mobile.

For D2I links, the control of mobility is split in LTE. A typical handoff is controlled at the RAN. In cases where anchor points (S-GW, P-GW) in the core network change as a result of the handoff, the MME, S-GW, and potentially P-GW can be involved to maintain the overlay routes (e.g., GTP tunnels). The selection of IP routes between the e-NB, S-GW, and P-GW are controlled via IP routing independent from mobility management. Furthermore, the MSPs have no choice but to use the control functionalities for these operations provided to them in their specialized hardware.

For D2D links, no mobility control operations have been defined in LTE yet. However, seamless connectivity needs to be maintained when the control plane for one or both of the UEs goes through a handoff, or when the D2D link is no longer feasible and has to be switched to a pair of D2I links.

In the proposed network architecture, the MSPs will have the option to conduct handoff and route management together and deploy different control applications for this purpose for different flows or users. We refer to this new paradigm as connectivity management as a service (CMaaS).

Connectivity Management as a Service

Legacy 3G/4G systems carry on the notion of providing almost lossless and low-delay handoffs between neighboring BSs. Such strict handoff management might be desirable for a high-quality (paid) VoIP service, but requiring mobile operators to provide the same stringent delivery performance for all flow types and non-paying OTT services is an expensive proposal as the connection and mobility management is pushed deeper into the telco cloud for denser small cell deployments. With the wide adaptation of new protocols such as DASH and using upper layer solutions that already have intelligence in maintaining session connectivity using new transport layer solutions (e.g., multipath TCP) or cloud based connection management, most signaling overhead due to handoff management may be offloaded to third party services for applications without stringent delay constraints. The proposed SDN-based 5G architecture, via the deployment of CPFs at different controller hierarchies, allows service differentiation at the level of connectivity management in a programmatic way for some network flows to achieve much higher performance as their service grades increase. This in turn results in the CMaaS offering and may be utilized by the MSP in one of three ways:

1 For services operated by the MSP itself, a connectivity management CPF that has a higher operational cost is used only for flows that require the associated stringent delay and packet loss levels. For others, an operationally cheaper CPF alternative is invoked.
2 A paying user may always be served by the higher-grade connectivity management of CPF regardless of its flow type.
3 For services operated by OTTs, the higher-grade connectivity management CPF is invoked only for paying OTTs.

Thus, CMaaS allows the MSP to lower its operational cost without sacrificing the quality of service for its own applications while introducing new revenue paths.

Unified Programmable Handoff and Routing Control for D2I Links

The all-SDN 5G-network architecture allows unified control of handoff and routing by jointly using controllers in the same or different hierarchies.

One possible realization of such unified control may be conducted at the network controller. In this case, for flows that are to be controlled by this grade level, each BS controller sends relevant feedback in the form of observed channel quality levels to the network controller, which in turn decides when a handoff occurs. This decision is pushed down to the BS controllers and simultaneously triggers the associated route update between the UE's new position and the egress node. For delay-sensitive flows, such a realization may not be desirable. However, handoffs intended for load balancing and/or user mobility handling may benefit from such a centralized approach. An MSP may decide to turn off a given BS to save energy and handoff all its users to neighboring BSs using this approach. Additionally, handoff decisions that might result in significant congestion in a given part of the CN may be avoided thanks to such a unified approach.

One alternative realization is to keep the handoff control at the BS and RAN controllers but in close coordination with the routing control that is realized at the network controller. In this case, a handoff triggers a routing and location tracking update. Depending on the desired service grade, this update may be reactive or proactive. When a reactive update is used, the network controller establishes a new route for the flow only after a handoff occurs. The flows in both directions are paused until the new route is established. Alternatively, using a proactive update, every time a handoff occurs, the network controller formulates a priori routing decisions for a given flow for all candidate next handoff locations, which may be made intelligently using a user's current mobility path and/or long-term mobility behavior to the egress node. This procedure should also be invoked at the setup of a new flow. Depending on whether there is a direct link between the source and target eNBs, the proactive handoff may involve either only the corresponding eNB controllers , additionally, the RAN controller. For the reverse path, the flow needs to be multicast to all candidate BSs to ensure lossless mobility management as described in [10]. Representative message flows for the proactive and reactive mobility control operations are given in Fig. 3. The execution of either of these approaches as a CMaaS requires extensions to OpenFlow. Details of such an execution are discussed in the next subsection after discussing how handoff control is accomplished for D2D links in the proposed architecture.

In the proposed network architecture, the MSPs will have the option to conduct handoff and route management together and deploy different control applications for this purpose for different flows or users. We refer to this new paradigm as connectivity management as a service.

Programmable Handoff Control for D2D Links

A D2D link has its data plane between the two UEs. However, the control plane of the link is between a controlling eNB (or a cluster head UE) and the UEs. Thus, the UE mobility potentially affects one or both of the control planes and/or the data plane. In this regard, as illustrated in Fig. 4, three types of handoffs may take place for a D2D link:

- Single mobility: The control plane of one of the UEs may be handed off to a different eNB.
- Dual mobility: The control planes of both of the UEs may be handed off to one or two new eNBs.
- D2D to D2I switching: The D2D link may become unsuitable for communication; thus, the data plane needs to be handed off to two D2I links, one per UE.

The proposed all-SDN network control architecture supports all three types of D2D handoffs.

To aid in the handoff decisions, the D2D link as well as the control plane link states need to be regularly fed back by the UEs to the controlling eNB [14]. We argue that for all three cases, only reactive handoffs are feasible as a proactive handoff control would result in very inefficient use of the wireless resources. The handoff procedure for the first two cases involves the UE controllers as well as the source and destination BS controllers. When the source eNB decides on a control plane handoff, it forwards this decision to the RAN controller, which in turn coordinates the resource management for the two eNBs for the D2D link. In the case of dual mobility, where the destination control plane eNB is the same for both UEs, the handoff control is handled by the two involved BS controllers, and the RAN controller is not included since no intercell coordinated resource management is necessary to sustain the D2D link. The D2D to D2I switching may potentially include the RAN and network controllers as well as the involved BS controllers, depending on how the two eNBs are to be linked to form the new route between the UEs.

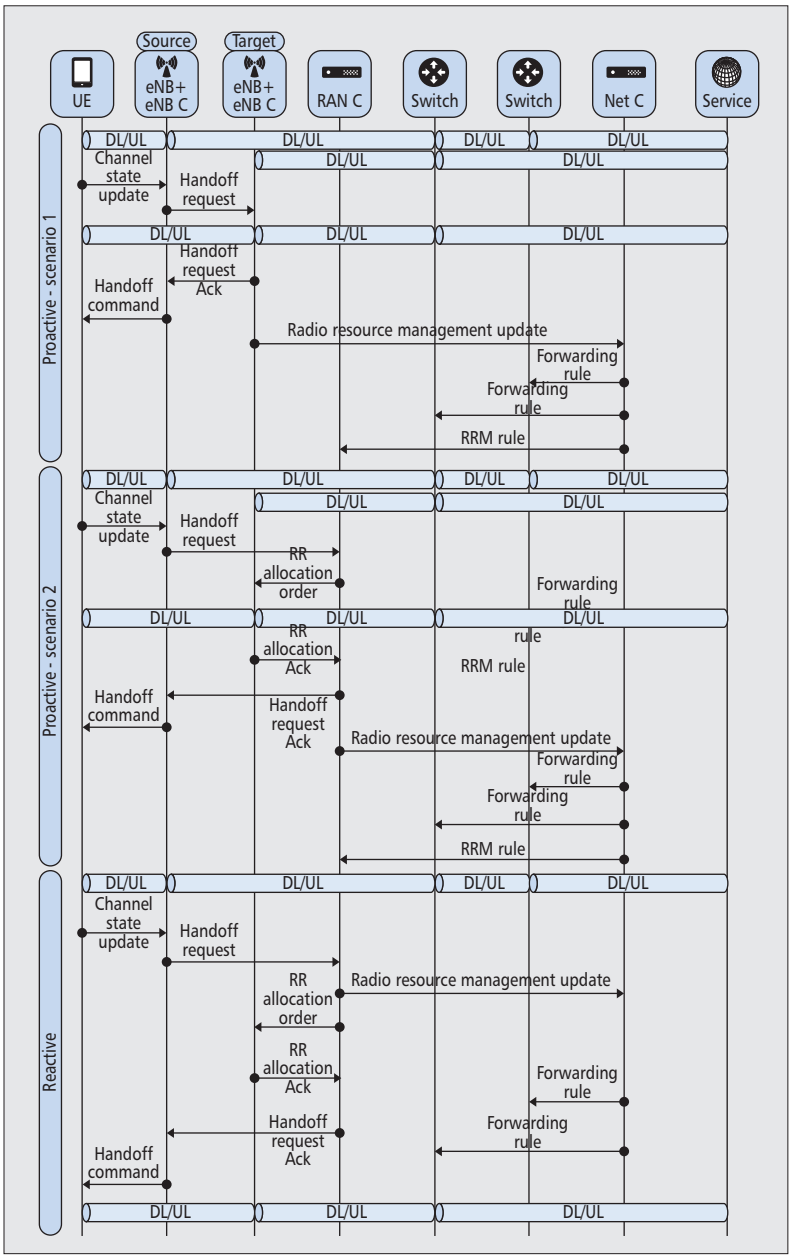


Figure 3. Message flows for proactive and reactive mobility management control of D2I links.

Performance Comparison of Reactive and Proactive CMaaS

We now investigate the performances of reactive and proactive CMaaS for D2I links, and reactive CMaaS for the D2D links in the proposed architecture to highlight the interplay between the data plane performance and control plane complexity in the proposed architecture. We investigate D2I and D2D cases separately, and consider a simple cellular network map with 1 GHz links, as illustrated in Fig. 5 for a geographical area of 10 eNBs. We investigate CMaaS at three different controller hierarchies using Mininet 2.1.0 and the Floodlight controller, generating rules that expire in 4 s. In the three scenarios, the RAN and network controllers are assumed to govern geographical areas of 10 eNBs and 7 switches, 25 eNBs and 14 switches, and 50 eNBs and 27 switches, respectively. The corresponding user populations for the three scenarios are assumed to be 100, 250, and 500. For the D2I mobility experiment, we assume that UEs are uniformly distributed across the eNBs at the beginning of the experiments. For the D2D mobility experiment, we further pair the UEs within each eNB to form direct links. Each UE is assumed to generate TCP flows randomly to one of the two available services for the D2I experiment, and to its paired UE for the D2D experiment, respectively, following an exponential distribution with parameter 1, where the flow durations are uniformly distributed between 0 and 10 s. Following this model, a UE may generate multiple parallel flows for a given duration of time. The sources in the network are assumed to generate flows with average data rates of 66 Mb/s. Each UE is assumed to go through a handoff every second following a random walk.

One hundred 30-min experiments have been conducted for reactive and proactive CMaaS in

this setup. For the D2I experiment, in the reactive mode, after a handoff occurrs, the controller is queried, a new route is computed toward the corresponding service, and the necessary set of rules are pushed down to all relevant network nodes. The active TCP flow is paused until the new route is established. In the proactive mode, the routes for all possible candidate handoff nodes are computed a priori so that routes are ready prior to the next handoff. In the experiments we assume that each eNB has six neighbors that are candidates for the next handoff. Even though the proactive mode prepares routes for the next handoff in advance, flows may still experience slight delays due to control plane operations. This is due to the 4-s expiry of each rule and the fact that the controller does not retransmit an already active rule to a switch and eNB, resulting in a slight possibility that the proactive CMaaS may contain reactive components.

The D2D experiment considers three possible handoff scenarios. For a given D2D link, depending on the mobility pattern, single-mobility handoff, dual-mobility , or D2D-to-D2I handoff may occur. Only the reactive mode is considered for D2D mobility. When single-mobility handoff occurs, the resultant setup is a D2D link that is controlled by two different eNBs. We assume that this handoff will result in a change of resource allocation for the D2D data link. The coordinated allocation of the resources for the two eNBs is managed by the RAN controller in this case. When dual mobility is encountered, the control planes of the UEs may both move to the same destination eNB or to different eNBs. If the move is to the same eNB, only the eNB controllers are involved in the handoff operation. Otherwise, the RAN controller is once again involved for coordinated resource allocation. We assume that for half the cases that result in dual eNB control of the D2D link, a handoff to D2I will be necessary. We assume in this case that the RAN and network controllers are involved in the handoff control along with the eNB controllers. The results for both D2I and D2D experiments are tabulated in Table 1.

For a D2I link, the current LTE handoff operation is a make-before-break scheme. On the other hand, the reactive CMaaS is a break-before-make scheme, whereas the proactive CMaaS is a make-before-handoff-request scheme. As such, the current LTE operation sits somewhere between the reactive and proactive modes. Specifically in LTE, once the source eNB decides to make a handoff, it continues to serve the UE until the end-to-end route for the UE via the target UE is established. However, during handoff, this link will be of low quality. Using the same network topology and UE mobility pattern, we conduct experiments to assess the LTE handoff performance. We assume an average 100 ms handoff operation duration [14], during which the link via the new eNB is to be established. During this time, we assume a 40 percent drop in the wireless channel capacity between the source eNB and the UE [4]. This is incorporated into the experiment by assuming that the wireless link allows only 60 percent of the full-quality link TCP throughput during the 100 ms of the handoff operation. We measure a corresponding average RTT delay of 2.01 ms and a corresponding throughput of 60.89 Mb/s for the TCP links.

For D2I handoffs, we observe that proactive CMaaS achieves round-trip time (RTT) delays that are less than 3 ms for all controller hierarchies, surpassing the performances of the LTE handoff procedure as well as the reactive CMaaS. The RTT delay is only 1.38 ms when there are 10 eNBs per network controller in the network. Reactive CMaaS, on the other hand, experiences larger RTT delays, increasing with higher controller hierarchy. Proactive CMaaS observes minimal loss in TCP throughput due to

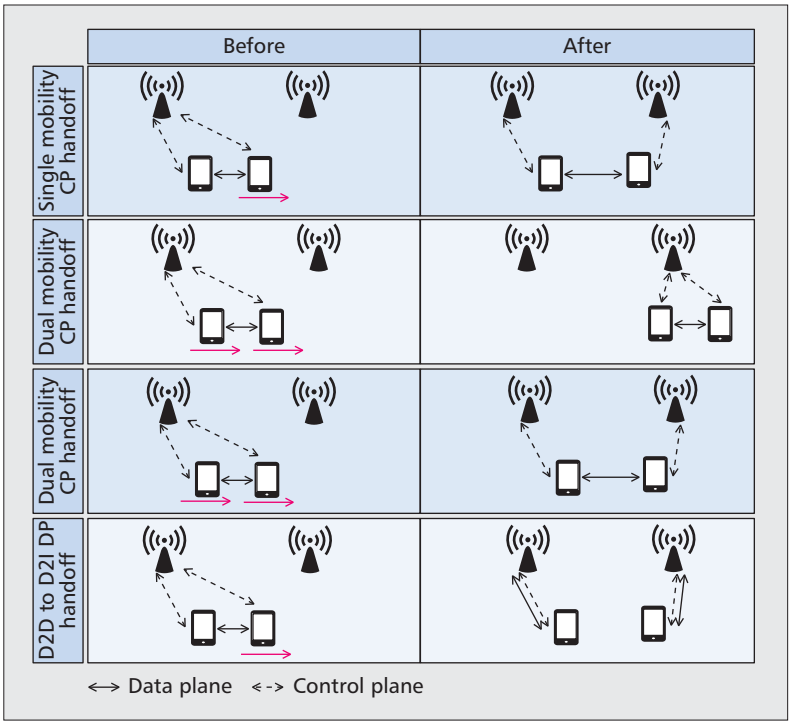


Figure 4. D2D mobility scenarios.

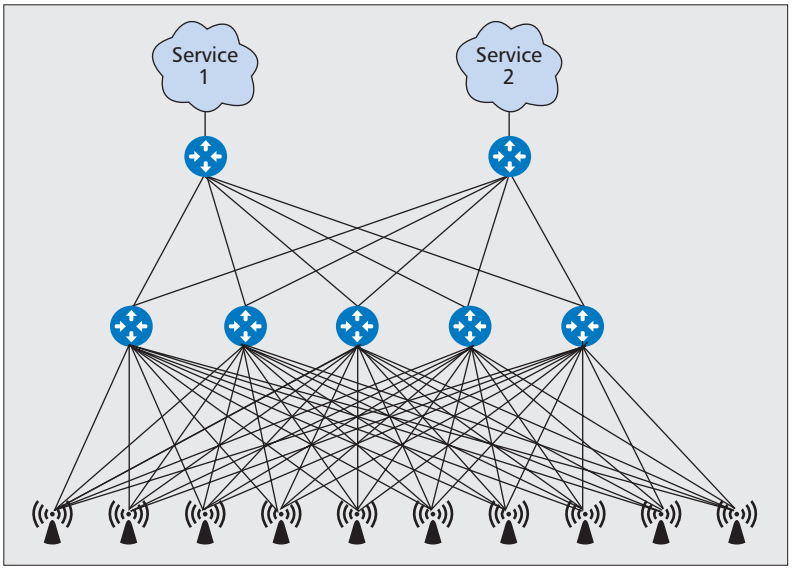


Figure 5. Cellular network map for CMaaS performance simulations.

The current LTE scheme provides a performance that lies between those of proactive and reactive CMaaS. The programmable nature of the proposed architecture allows for any of the three (or other) schemes to be deployed where this decision may be made on a per flow basis.

Proposed all-SDN 5G control plane architecture						
D2I link mobility						
Network controller coverage						
	10 eNBs and 100 UEs		**25 eNBs and 250 UEs**		**50 eNBs and 500 UEs**	
Performance criteria	**Reactive CMaaS**	**Proactive CMaaS**	**Reactive CMaaS**	**Proactive CMaaS**	**Reactive CMaaS**	**Proactive CMaaS**
Average TCP flow RTT delay	6.46 ms	1.38 ms	10.55 ms	1.89 ms	17.25 ms	2.73 ms
Average TCP flow throughput	62.64 Mb/s	63.42 Mb/s	62.01 Mb/s	63.34 Mb/s	60.98 Mb/s	63.21 Mb/s
Average network controller rules per second	2148.2	2686.0	5420.2	7512.4	10,838.8	15,512.8
D2D link mobility						
RAN controller coverage						
	10 eNBs and 100 UEs		**25 eNBs and 250 UEs**		**50 eNBs and 500 UEs**	
Average TCP flow RTT delay	2.70 ms		2.86 ms		3.05 ms	
Average TCP flow throughput	63.22 Mb/s		63.19 Mb/s		63.16 Mb/s	
Average RAN controller rules per second	181.3		477.4		970.8	
Average eNB controller rules per second	20.0		20.0		20.0	

Table 1. Reactive and proactive CMaaS performance for D2I and D2D links with SDN controllers controlling different geographical areas.

handoffs, while the corresponding loss is higher in reactive CMaaS. The superior data plane performance of proactive CMaaS comes at the expense of an increased complexity control plane, as shown by the larger number of control plane rules that need to be computed and communicated by the controller every second compared to reactive CMaaS. As expected, the current LTE scheme provides performance that lies between those of proactive and reactive CMaaS. The programmable nature of the proposed architecture allows for any of the three (or other) schemes to be deployed where this decision may be made on a per flow basis. It should be noted here that today's SDN controllers are capable of responding up to 1,000,000 flows/s when run on Amazon's Elastic Computer Cloud using a Cluster Compute Eight Extra Large instance, containing 16 physical cores from 2 × Intel Xeon E5-2670 processors, 60.5 Gbytes of RAM, using a 64-bit Ubuntu 11.10 VM image [15]. Despite larger incurred delays, reactive CMaaS is still useful for an MSP for deployment of delay-tolerant services, especially when control plane capacity is critically needed for some other application in the network or some of the eNBs are critically loaded.

For D2D handoffs, we observe that even though only a reactive approach is considered, the observed RTT delays are very small. Indeed, the delay is only 2.7 ms when there are 10 eNBs for each RAN controller. We observe that the TCP flow throughputs of the mobile D2D links are comparable to those of the mobile D2I links.

Conclusions

The wireless Internet is experiencing tremendous growth thanks to the introduction of smart phones and associated bandwidth-hungry applications. The MSPs are struggling to keep up with this demand in today's networks. The 5G architecture needs to bring a high-capacity, agile, low-cost solution to ensure both user satisfaction and MSP profitability. This article introduces a programmable all-SDN architecture with hierarchical network control capabilities to allow different grades of performance for all fundamental network functionalities. Using this architecture, we then introduce a unified approach to mobility, handoff, and routing management proposal, connectivity management as a service. CMaaS allows the control of mobility and routing of different flows or users differently in the network,

thereby opening a new revenue generation path to MSPs.

References

[1] ETSI, "Network Functions Virtualization (NFV): Use Cases," ETSI Group Spec. NFV 001 v1.1.1, Oct. 2013.
[2] K. Chen and R. Duan, "C-RAN: The Road Towards Green RAN," China Mobile Research Institute. White Paper, Oct. 2011.
[3] J. Kempf *et al.*, "Moving the Mobile Evolved Packet Core to the Cloud," *2012 IEEE 8th Int'l. Conf. Wireless and Mobile Computing, Networking and Commun.*, 2012, pp. 784–91.
[4] H. Holma and A. Toskala, Eds. *LTE for UMTS – Evolution to LTE-Advanced*, 2nd ed., Wiley, 2011.
[5] 3GPP, "Overview of 3GPP Release 12," v. 0.1.3, June 2014.
[6] X. Lin *et al.*, "An Overview of 3GPP Device-to-Device Proximity Services," *IEEE Commun. Mag.*, vol. 52, no. 4, Apr. 2014, pp. 40–48.
[7] ONF, "OpenFlow Switch Specification," v. 1.4.0, Oct. 14, 2013.
[8] K. Pentikousis, Y. Wang, and W. Hu, "MobileFlow: Towards Software Defined Mobile Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013, pp. 44–53.
[9] X. Jin, L. Li, and J. Rexford, "SoftCell: Scalable and Flexible Cellular Core Network Architecture," *Proc. ACM CoNEX)*, ACM, Dec. 2013.
[10] K. Yap *et al.*, "Blueprint for Introducing Innovation into Wireless Mobile Networks," *Proc. ACM SIGCOMM Wksp. Virtualized Infrastructure Sys. and Architectures*, Sept. 2010.
[11] M. Bansal *et al.*, "OpenRadio: A Programmable Wireless Dataplane," *Proc. ACM SIGCOMM Wksp. Hot Topics in Software Defined Networking*, 2012, pp. 109–14.
[12] A. Gudipati *et al.*, "SoftRAN: Software Defined Radio Access Network," *Proc. ACM SIGCOMM Wksp. Hot Topics in Software Defined Networking*, Aug. 2013.
[13] S. Kumar *et al.*, "Bringing Cross-Layer MIMO to Today's Wireless LANs," *Proc. ACM SIGCOMM 2013 Conf.*, 2013, pp. 387–98.
[14] A. Racz, A. Temesvary, and N. Reider, "Handover Performance in 3GPP Long Term Evolution (LTE) Systems," *16th IST Mobile and Wireless Communications Summit*, 2007.
[15] D. Erickson, "The Beacon OpenFlow Controller," *Proc. ACM SIGCOMM Wksp. Hot Topics in SDN*, 2013.

Biographies

VOLKAN YAZICI (volkan.yazici@ozu.edu.tr) is currently a Ph.D. student in the Department of Computer Engineering at Özyeğin University, Istanbul, Turkey. He received his M.Sc. degree in computer engineering from Bilkent University, Ankara, Turkey. He has participated in national and EU scientific projects. His research interests include combinatorial optimization, graph and hypergraph models, data structures and algorithms, Internet2 technologies, and software-defined networks.

ULAŞ C. KOZAT [SM] (kozat@docomoinnovations.com) is a principal researcher and project manager at DOCOMO Innovations Inc., Palo Alto, California. He also serves as an adjunct associate professor in the Electrical Engineering Department at Özyeğin University. He received his Ph.D., M.Sc., and B.Sc. degrees, all in electrical engineering, from the University of Maryland College Park, George Washington University, Washington, DC, and Bilkent University, respectively. He has been conducting research in the broad areas of networking and wireless communications.

M. OĞUZ SUNAY [SM] (oguz.sunay@ozyegin.edu.tr) is an associate professor of Electrical and Electronics Engineering at Özyeğin University. He received his Ph.D. and M.Sc. degrees from Queen's University, Kingston, Ontario, Canada, and his B.Sc. degree from ODTÜ, Ankara, Turkey. Prior to academia, he worked at Nokia Research Center, Texas, and Bell Laboratories, New Jersey, where he actively participated in 3GPP standardization efforts. He holds over 25 patents in various aspects of wireless communication systems. His current research areas lie in the areas of wireless communications and networks, with special focus on software-defined wireless networks, programmability and virtualization in wireless networks, and dynamic resource allocation.

This article introduces a programmable all-SDN architecture with hierarchical network control capabilities to allow different grades of performance for all fundamental network functionalities.

Terminal-Centric Distribution and Orchestration of IP Mobility for 5G Networks

Alper Yegin, Jungshin Park, Kisuk Kweon, and Jinsung Lee

ABSTRACT

Advances in radio access technologies and mobile terminals are fueling the growth of Internet traffic via mobile networks. The upcoming 5G era, with a new 5G radio technology and increased utilization of heterogeneous networking, will further accelerate mobile data usage. One of the pillars of mobile network architecture is IP mobility management, which is currently based on centralized data path management. In this article, we first describe the efficiency issues of the centralized approach. We then discuss multiple dimensions of distributing the mobility functions, and suggest how the mobile terminal can utilize them in orchestration for efficient communication over a 5G flat mobile network.

INTRODUCTION

Internet usage continues its rapid expansion thanks to the technological advances in wired and, more importantly, wireless access technologies, accentuated by a stream of increasingly capable mobile terminals entering the consumer market, and new services in communication, entertainment, commerce, and productivity areas boosting the utility of the Internet. While overall IP traffic is expected to have 23 percent annual growth between 2012 and 2017, IP traffic from mobile terminals is expected to have 66 percent annual growth during the same period [1]. This does not come as a surprise as we are witnessing mobile becoming the preferred access method for users; also, the Internet of Things is starting to roll out.

The upcoming traffic surge needs to first be addressed at the radio access level. There are already industry efforts underway [2] to design the next generation radio access technology to meet the traffic demand for 2020 and beyond, also known as the fifth generation (5G) era. While design of a new radio access technology is inevitable, it also needs to be accompanied with design of a new network architecture in order to tackle the problem from multiple fronts. We envision a new architecture that will work with both legacy and new 5G radio access technologies.

Providing a stable data path in the face of terminals changing their point of attachment to the network is the essential issue that will drive the new architecture design. In the legacy network architectures, a terminal's traffic is routed through a centralized node in the mobile core network. This centralized node acts as an anchor for the data path and ensures that IP packets reach the terminal irrespective of its point of attachment.

As IP traffic grows and increasingly migrates onto the mobile access networks, the inefficiencies of the centralized mobile architecture become more serious. The fixed/non-mobile Internet architecture is very efficient and scalable due to its distributed design. The next generation mobile network needs to adopt the Internet architecture design principles in order to match its scalability and efficiency characteristics. Specifically, the mobile networks need to adopt the distributed nature of IP routing. Providing mobile data path management on top of a distributed architecture is one of the fundamental challenges of 5G design.

In the following sections we first describe the issues in the current mobile network architectures. We then discuss the proposed IP mobility architecture design by showing its founding principles, and how they are integrated and used in coordination. Subsequently, we describe our proposed design's expected effect on the mobile networks, followed by a review of relevant industry activities.

LEGACY ARCHITECTURES

In the 3G/4G mobile network architectures a dedicated gateway in the core network (e.g., PGW — packet data network gateway in the Third Generation Partnership Project, 3GPP) acts as an IP anchor. The gateway allocates an IP address to the terminal, tracks the location of the terminal within the IP topology, and ensures the terminal's reachability by tunneling traffic to its points of attachment (Fig. 1).

In this design, data path management is trans-

The authors are with Samsung Electronics Co., Ltd.

0163-6804/14/$25.00 © 2014 IEEE

parent to the IP stack of the terminal. All IP flows of the terminal are subject to the same mobility treatment regardless of whether they need it or not.

According to this design, the end-to-end data path between the terminal and its communication peer follows a route that is forced to go via the mobile core network as opposed to following the most direct data path provided by native IP routing. This design favors simplicity while causing the following side effects:

- The end-to-end transmission delay is increased due to elongated data path.
  The end-to-end transmission delay can be reduced by approximately 30 percent [3] when the traffic is carried over the shortest data path as opposed to being transported on a long-haul triangular path via a core network.
- There is additional load of backhauling and network processing in the core network.
  Offloading general IP traffic from the core network yields cost savings due to reduction of backhauling and core network routing resources. Approximately 25–35 percent cost savings are foreseen for operators when offloading general IP traffic from their core network [4].
- Network reliability is reduced due to introduction of a single point of failure, the PGW [5].
  A single gateway or a single site hosting multiple gateways that is located in the core network and responsible for maintaining data path state and forwarding traffic for all terminals emerges as a single point of failure.

A distributed design that places the intelligence at the edges of the network is one of the fundamental principles of Internet architecture. Departure from this principle causes the aforementioned issues, which are also amplified as more and more IP traffic moves to mobile networks.

Furthermore, network operators are interested in placing content servers as close to the access network as possible in order to cut down on backhauling cost and transmission delay. However this is not possible in the current network architecture as IP traffic is routed through the core network before reaching any destination.

One of the simplest solutions to overcome the aforementioned issues in the current 3G/4G network architecture is to place the gateway in the radio access network. This approach can provide direct data paths, but also leads to increased IP session breakage as each handover is more likely to result in a gateway and IP address change. The selected IP traffic offload (SIPTO) feature in 3GPP is a good example of this approach that aims at providing direct data paths, but fails to address the mobility aspect.

A well established term for networks that allow direct IP connectivity from a gateway close to or co-located with base stations is network flattening. A typical Wi-Fi hotspot with its co-located access point and access gateway functionality and direct connection to the Internet (i.e., no tunneling to a core network) is an example of a flat network. The real challenge of next generation mobile networks is to have as flat a network as possible with full-fledged mobility support.

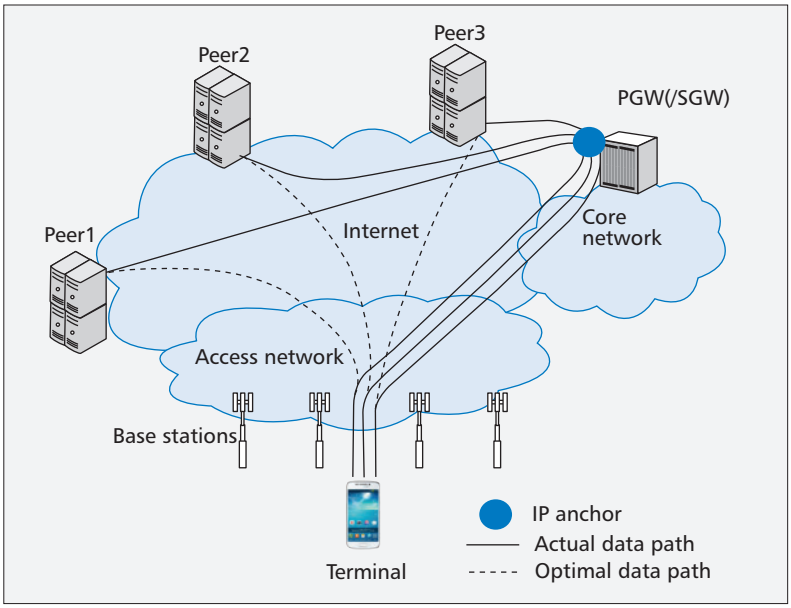


Figure 1. Centralized mobile data path management.

Multidimensional Distribution

The reasons a centralized design would not have brought success to the Internet are equally applicable to why it would not work for mobile Internet. Starting from that premise, we first identify a number of design approaches that focus on distribution of mobility functions. The identified design approaches are collectively utilized in our design with the help of terminal orchestration.

Distribution Across IP Anchors

The current 4G architectures (3GPP LTE and WiMAX) designate a dedicated IP gateway (PGW, HA — home agent) in the core network to be the sole IP anchor for the terminal. Accordingly, the terminal is assigned an IP address/prefix from the pool of IP prefixes managed by that gateway, and every IP packet in/out of the terminal is forced to traverse it. This gateway is considered to be performing core anchoring according to our location-based IP anchor classification.

In order to mitigate the aforementioned efficiency issues associated with the centralized approach, we propose using access anchoring and remote anchoring, both of which happen at the edges of the Internet.

A gateway placed in the access network (XGW) can allocate an IP address/prefix to the terminal from its own pool of prefixes, and ensure continued use of that IP address for ongoing IP sessions (e.g., a VoIP call, web page download) even after the terminal has moved under another gateway with the help of a forwarding tunnel between itself and either the terminal or the other gateway [6–8]. The anchoring is terminated as soon as the last flow using the anchored IP address terminates. This is called access anchoring.

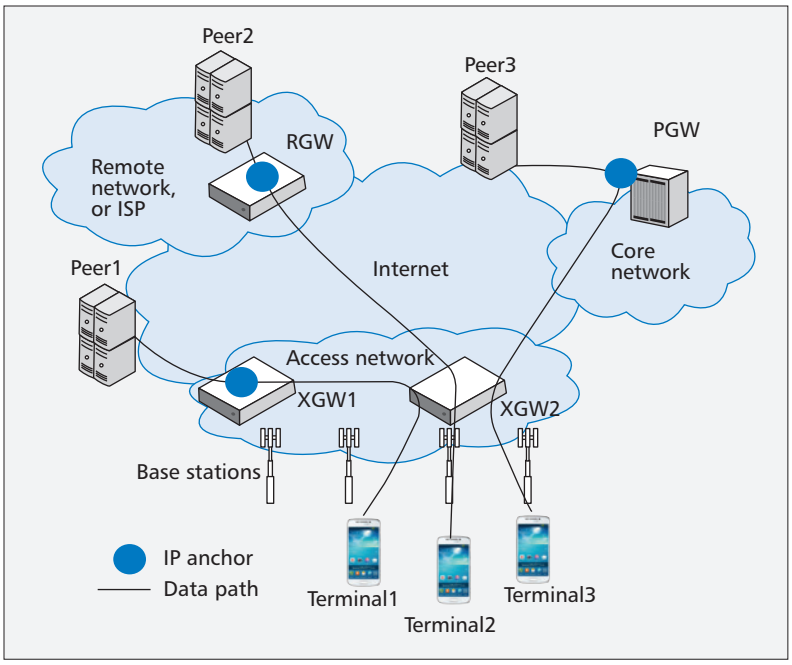


Figure 2. Access, remote, and core anchoring.

Access anchoring can support IP session continuity: the ability to provide an IP address to the terminal that remains valid for the entire duration of a flow. But it is not suitable to support fixed IP address allocation to the terminal, which requires a centrally located anchor for the terminal at all times as it may roam all around the world.

Access anchoring provides the optimal data path between the terminal and its communication peer (remote end) at the beginning of the communication. The data path starts to become suboptimal for long-standing IP flows when the terminal moves far away from the anchor. Note that even a small physical movement can yield a large topological jump when considering, for example, handovers between cellular and Wi-Fi networks operated by two distinct service providers.

Remote anchoring, on the other hand, aims to provide an optimal data path despite terminal movements. A remote anchoring gateway (RGW) is placed near the communication peer of the terminal, for example, within the peer network or Internet service provider (ISP) serving that network. Such a gateway assigns a session IP address to the terminal for its communication with the designated peer. It acts as a mobility anchor by keeping track of terminal handovers and delivering the IP packets destined for the assigned IP address to the terminal's topological IP address at its current attachment [9]. It is expected that this type of anchor would be deployed by large content sites (e.g., Facebook, YouTube, Gmail) or the ISPs and/or CDN providers serving such content sites based on a new business model. Use of remote anchoring enables the mobile operator to offload the mobility task and the associated traffic to such third parties. Because an RGW is close to the peer, it is also guaranteed to stay on/near the optimal data path regardless of a terminal's location and movement.

Remote anchoring only supports IP session continuity. It cannot support fixed IP address allocation as the anchored IP address is dynamically assigned to the terminals. While remote anchoring achieves a shorter data path than access anchoring, the latter achieves better seamless handovers because of using short-haul signaling. The terminal can achieve both data path optimization and seamless handovers by using remote anchoring throughout the IP session, and using access anchoring transiently during IP handovers until the RGW is updated with the terminal's new topological IP address.

Figure 2 depicts the use of three different types of IP anchoring by terminals.

In our proposed network architecture, we replace the responsibility of a single and fixed IP anchor in the core network with a number of anchors scattered across the edges of the Internet. These anchors are dynamically engaged and released in order to optimize the data paths.

Distribution across Flows

IP flows differ with respect to their mobility management needs. Flows used by a server application running on the terminal (e.g., a mobile camera application serving remote clients) require a fixed IP address on its local end so that the incoming connections can find the server application at a known/published IP address. On the other hand, typical client applications do not need a fixed IP address, as they are the initiator of the communication. They can choose any available IP address as the source address for communication. However, some of these client applications (e.g., live video streaming) require that the IP address does not change during an ongoing IP session for session continuity. Furthermore, some flows need neither a fixed IP address nor IP session continuity. These may be short-lived flows that can complete their task in a few quick round-trips (e.g., DNS queries), flows capable of achieving session continuity with the help of higher-layer mobility protocols (e.g., SIP-based IMS), or flows of applications with built-in methods to handle IP address changes (e.g., video clients reconnecting to the video server upon IP handover while smoothing the transition with the help of a video download buffer).

Flows requiring fixed IP addresses need to be served by core anchoring. Flows requiring IP session continuity can be served by either remote or access anchoring. Flows requiring neither of these features can survive without any IP anchoring.

A terminal may be running a potpourri of flows with respect to their types. Treating all flows the same way, and providing both fixed IP address and IP session continuity indiscriminately to each of them by means of core anchoring is overkill. Instead, the terminal stack should be able to distinguish flows with respect to their mobility needs and treat each of them individually. The network stack can rely on implicit (e.g., profiling based on application identifier, destination hostname/IP address, port number of the flow) and application-provided explicit indications [10] to identify the flow's needs.

Ideally, using an unanchored IP address is

the most efficient way of using network resources. The terminal can start operating with an unanchored IP address, and dynamically configure an anchored IP address as soon as it receives a flow initiation request that implicitly/explicitly indicates a need for anchoring. The dynamically configured anchored address can be released when there are no more flows using that address. Thus, IP addresses get added and deleted dynamically in response to the types of flows that are created and terminated (Fig. 3). The source IP address selection mechanism on the terminal needs to be augmented to take mobility needs into account when binding a socket to one of the available IP addresses based on its mobility (i.e., anchoring) characteristics, and to trigger dynamic configuration of additional IP addresses when the required type of address is not available on the terminal.

Figure 3 depicts an example where Flow1 requires IP session continuity, and hence is bound to IP address IP1, which is assigned by/anchored on XGW1 (shown after handover to XGW2), Flow2 requires no mobility support, and hence is bound to IP2, which is assigned by XGW2 (the serving gateway, no anchoring), and Flow3 requires a fixed IP address, and hence is bound to IP3, which is assigned by the PGW.

In summary, applying mobility on a per-flow basis as opposed to a monolithic treatment across the terminal is proposed as another dimension of mobility management distribution. It should be noted that less granular approaches, such as mobility on a per-terminal or per-user basis, are also possible.

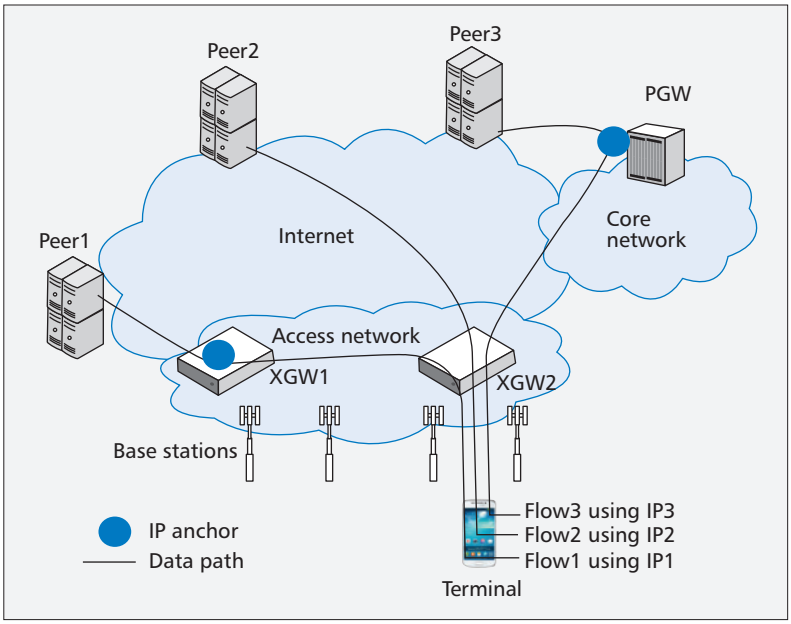


Figure 3. Mobility treatment on a per-flow basis.

Distribution across Communication Layers

IP session continuity is not the only way to achieve application session continuity [11]. For example, sessions established using SIP or MPTCP can survive IP address changes using the mobility management support built into those protocols. Furthermore, some applications (e.g., instant messengers) can deal with IP address changes on their own. Such applications detect a terminal's IP address change and notify their peers of the new IP address (e.g., Skype).

The advantage of using a layer 4 or above (L4+) solution is its ability to set up an optimal data path between the endpoints by using their topological IP addresses [12, 13]. In that case the IP packets are not forced to traverse an off-path central gateway because of using a non-topological source or destination address.

Availability of the same L4+ solution on both endpoints of a communication cannot be guaranteed because of the heterogeneous nature of the Internet. Furthermore, even though L4+ solutions can achieve data path optimization, they perform poorly during handovers due to their end-to-end nature. The required data path update signaling at each IP handover needs to traverse the Internet, and leads to loss of in-flight packets during the end-to-end state convergence.

On the other hand, IP layer (Mobile IP) and sub-IP layer solutions (Proxy Mobile IP, GTP) are good at dealing with seamless handovers, because data path extensions between and within access networks (e.g., an inter-XGW tunnel) can be set up in response to terminal mobility. In addition, their deployment does not rely on any support from the terminal's communication peer. However, they cannot set up the optimal data path as they cause triangular routes.

Besides the trade-off between the IP/sub-IP and L4+ solutions, there is also undesirable interaction between them. When an IP/sub-IP solution is running, it obstructs operation of any L4+ solution. The former type ensures that the IP address seen by the higher layers stays fixed in spite of the terminal changing its location within the IP topology. This special effect makes the L4+ solutions think that a terminal is stationary; hence, they do not attempt to update their peers. As a result the terminal is subject to suboptimal data paths despite the presence of an L4+ solution that can remedy the problem.

Given that there is a trade-off and an undesirable side effect between the two types of solutions, we find that the best result can be achieved when these solutions are used in orchestration: the terminal shall prefer L4+ solutions as the main actor for handling mobility and augment their performance with transient use of IP/sub-IP solutions within the access network scope.

In our architecture, the type of mobility protocol to use is determined on a per-flow basis. For example, the terminal would rely on MPTCP mobility only if the application is using TCP, and MPTCP is supported by the terminal and its communication peer. Given that not all nodes across the Internet implement and use the same or all available L4+ protocols, it is likely that flows on a given terminal may be subject to different mobility solutions, with some having to fall back to relying solely on IP/sub-IP solutions (Fig. 4).

IP/sub-IP mobility solutions shall only be transiently applied to flows that are mainly relying on L4+ mobility solutions. When the termi-

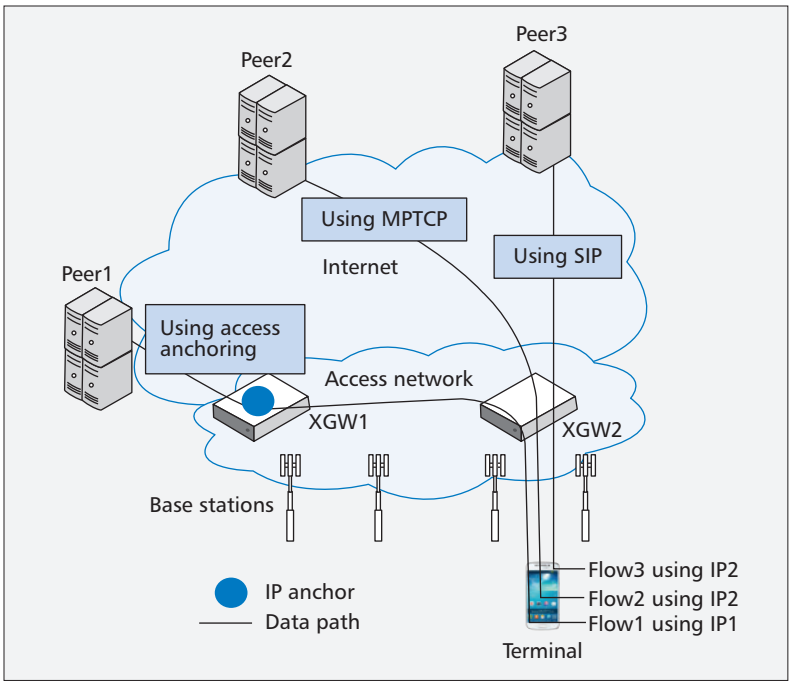


Figure 4. Each flow using a mobility solution at a different communication layer.

nal moves from one access gateway to another, if the gateways are supporting IP/sub-IP mobility, a forwarding tunnel is set up between the anchor gateway and either the serving gateway or terminal. This ensures continuation of an IP address on the terminal until the terminal has successfully updated its peer with its new IP address using the L4+ solution. The transient tunnel is not needed after the terminal has successfully updated its peer with its new IP address using the L4+ solution.

This dimension of mobility distribution suggests that optimal results can be achieved when session continuity is handled by orchestrated execution of multiple protocols at different layers.

Distribution between Network and Terminal

Traditionally, IP mobility has been under full control of the mobile networks and transparent to the terminals. According to the aforementioned mobility distribution approaches, it is clear that the terminal needs to be actively involved in next-generation mobility management.

The terminal is in the best position to identify and characterize its own flows, so it can perform per-flow mobility management. In contrast, the network infrastructure cannot match the same capability even by using DPI methods, which appears to be a costly solution with limited capabilities (application meta-data is only available to the terminal stack, and an anchoring decision needs to be made even before the network sees the very first packet of a new flow).

Furthermore, only the terminal is aware of the availability of third party networks within its reach (e.g., free Wi-Fi at home or in the office). These third party networks do not have signaling interfaces with the operator networks; therefore, no sub-IP layer solution can work in between. Only IP and higher-layer solutions can achieve session continuity in such networks, and they require cooperation from the terminal. Increased use of third party networks is beneficial for both the user and the operator for cost reduction. Such a benefit would be maximized when the terminals can perform seamless handover across these networks.

Finally, L4+ solutions operate end to end, and therefore require a terminal's active engagement in signaling. Intermediate networks are not part of the end-to-end signaling, such as MPTCP, barring any hacks.

For these reasons, it is expected that the main IP mobility function would be hosted on the terminal. It would decide IP mobility treatment of each flow, and allow the terminal to orchestrate available mobility solutions to serve the flow. Meanwhile, the network would still be in charge of handling the sub-IP solutions.

Orchestration

Orthogonal design principles come together to form the basis of our 5G mobility management design. The terminal is at the center of this design, in charge of orchestrating the mobility management execution [14].

The foundation of the design is built on the fact that mobility treatment applies on a per-flow basis, not on a per-terminal basis. The architecture allows each flow to be treated individually based on its mobility needs. Figure 5 depicts how the terminal stack determines the main mobility solution to apply on a given flow.

The network stack on the terminal identifies the mobility need of a flow either based on the associated application's explicit indication via an API [10] or by implicit profiling. If the flow is determined not to require any special mobility treatment, no mobility support is assigned to the flow. If the flow requires a fixed IP address, core anchoring is assigned to it.

An L4+ mobility solution is the preferred choice when a flow requires session continuity, and both endpoints support at least one common L4+ mobility protocol that is applicable to the flow. In the absence of such a solution, the terminal stack attempts to fall back to using remote anchoring if available. Remote anchoring is preferred over access anchoring as the former does not create a triangular data path. Even access anchoring may not be available considering the use of unmanaged third party networks. In that worst case scenario, the terminal falls back to using core anchoring for IP session continuity support as well.

Each selected mobility method results in binding the flow to a matching IP address configured on the terminal. When mobility support is not required or an L4+ solution is selected, the flow is bound to an unanchored IP address provided by the nearest IP gateway. IP addresses bound to flows using core, remote, and access anchoring are provided by gateways located in the core, remote, and access networks, respectively.

The terminal stack may or may not already have the required type of IP address configured when the application attempts to initiate a new flow. If the required type IP address is not present, it gets dynamically configured and bound to the flow. A dynamically configured IP address can be freed after the last flow using the address has terminated, either immediately or after a grace period. Therefore, the terminal is expected to have a dynamic set of IP addresses at any point in time, and to fall back to a single unanchored IP address when there is no flow with any special needs.

Access anchoring acts as a supplement when an L4+ solution or core/remote anchoring is used. A temporary forwarding tunnel is set up between the anchor (previous) gateway and either the terminal or serving gateway in order to allow the terminal to keep using its data path through the previous gateway until its peer or the remote anchors are updated with the new IP address of the terminal.

Network Flattening

Our design aims at restoring the routing efficiency and reliability attributes of the mobile networks that are currently jeopardized due to the data path distortion caused by centralized mobility management.

L4+ mobility solutions ensure that the data packets follow the shortest path suggested by the native IP routing. For example, video download/streaming clients can seamlessly handle mobility at the application layer by performing automatic reconnections to the server upon IP handovers and using video download buffers. Our design favors utilization of such solutions and ensures their unobstructed performance. This approach promotes increased deployment of L4+ mobility solutions and design of new ones.

Remote anchoring is another solution that can achieve data path optimization. Due to high concentration of mobile traffic in major content sites, application of this solution to a few content providers can reduce the transmission delay for a significant part of user traffic (e.g., YouTube, Facebook, Google Market, Pandora, and Netflix collectively generate 50 percent of all mobile Internet traffic in the United States today [15]).

In current architectures the majority of the IP flows' reliance on core anchoring is simply because of its mobility support. They would rather use an L4+, remote anchoring, or access anchoring solution when these solutions are available. The only types of flow that still need to traverse the core network are special ones such as mobile router or mobile server flows, or 3GPP-style enterprise VPN flows (using a dedicated packet data network connection for accessing the enterprise network). These flows require either a fixed IP address or an IP address from a fixed address pool that is managed by a gateway in the core network. Additionally, flows subject to lawful intercept are also forced to go via a core network due to regulations. These flows constitute a negligible fraction of the overall mobile traffic.

An obvious side effect of our design is the increased IP address consumption. Continued use of private IPv4 addresses and eventual transition to global IPv6 addresses would handle the IP address needs of our design. Furthermore, use of IP/sub-IP solutions for achieving seamless handovers along with L4+ solutions introduces additional bidirectional signalling between the previous gateway and the terminal or serving gateway at each IP handover.

There have been several attempts to flatten mobile networks in past generations, but so far they have achieved only small improvements as shown by their centralized design. However, we expect that true flat network architecture can be realized by our design proposal in order to cope with the challenging demands of the 5G era.

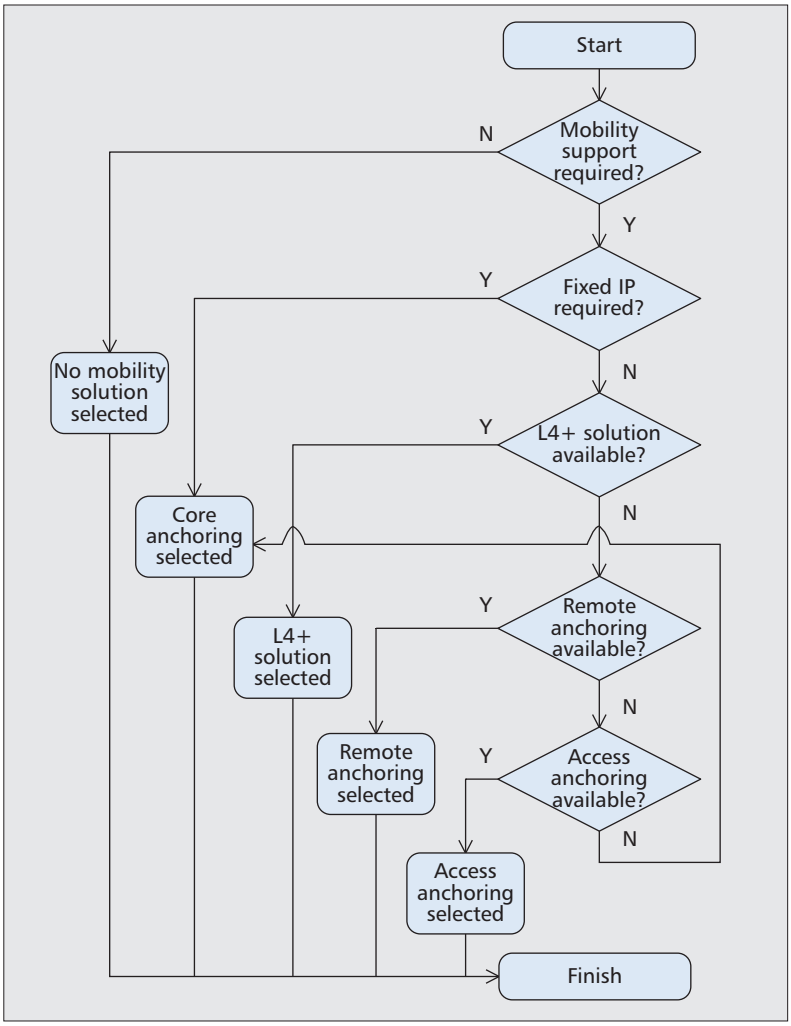


Figure 5. Algorithm for determining the main mobility solution for a given flow.

Related Standards Activities

There are currently two ongoing standards activities related to our vision: the Internet Engineering Task Force (IETF) Distributed Mobility Management (DMM) Working Group and 3GPP Coordinated PGW Change for Selected IP Traffic Offload (CSIPTO) Work Item. The scope of the former is distribution of IP anchors toward the network edge, and the latter is about termi-

The current and past generations of mobile networks use a centralized approach for managing mobile data paths. This approach has served the networks well as long as they carried small amounts of IP traffic. We propose replacing the centralized mobility approach with a distributed approach that places the intelligence on the mobile terminal in order to achieve scalability and efficiency in 5G networks.

nal assisted access anchoring and facilitation of L4+ mobility.

Currently these activities are limited in scope and lack any coordination. We consider them the first steps toward realizing our vision. It is expected that follow-up work items will be launched in the respective organizations and proceed in coordination for setting the global standards in time for 5G networks.

Conclusion

The current and past generations of mobile networks use a centralized approach for managing mobile data paths. This approach has served the networks well as long as they carried small amounts of IP traffic. Mobile access is already increasing its stake in the overall IP transport, and that is bound to accelerate with the introduction of 5G radio access technology. We propose replacing the centralized mobility approach with a distributed approach that places the intelligence on the mobile terminal in order to achieve scalability and efficiency in 5G networks. We reveal multiple dimensions of functional distribution for optimal mobile data path management and describe their orchestration by the mobile terminal. We also identify the early signs of industry standards starting to move in this direction.

Acknowledgments

The authors would like to thank Joohyung Lee and Antony Franklin (Samsung Electronics), Avi Lior (Rockport Networks), and the anonymous reviewers for their helpful comments, which significantly improved the quality of this article.

References

[1] White paper, "Cisco Visual Networking Index: Forecast and Methodology, 2012–2017," Cisco VNI Report, May 2013.

[2] W. Roh *et al.*, "Millimeter-Wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014.

[3] D. Liu, J. Song, and W. Luo, "Centralized and Distributed Mobility Traffic Analysis," IETF 80 MEXT Working Group, Mar. 2011.

[4] White paper, "Top 10 Considerations for a Successful 4G LTE Evolved Packet Core Deployment," Cisco C11-730105-00, 2013.

[5] P. Bosch *et al.*,"Flat Cellular (UMTS) Networks," *IEEE WCNC 2007*, Hong Kong, Mar. 2007.

[6] P. Bertin, S. Bonjour, and J.-M. Bonnin, "An Evaluation of Dynamic Mobility Anchoring," *Proc. IEEE VTC-Fall '09*, 20–23 Sept. 2009, Anchorage, AK, 2009, pp. 1–5.

[7] T. Condeixa and S. Sargento, "Dynamic Mobile IP Anchoring," *Proc. 2013 IEEE ICC*, 9–13 June 2013, Budapest, Hungary, pp. 3607–12.

[8] J. Lee *et al.*, "Mobile Data Offloading: A Host-Based Distributed Mobility Management Approach," *IEEE Internet Computing*, vol. 18, no. 1, Jan.-Feb. 2014, pp. 20–29.

[9] A. Yegin *et al.*, "Corresponding Network Homing," IETF Internet draft, draft-yegin-dmm-cnet-homing-01, Oct. 2013.

[10] A. Yegin *et al.*, "On Demand Mobility Management," IETF Internet draft, draft-yegin-dmm-ondemand-mobility-01, July 2013.

[11] E. Perera *et al.*, "A Mobility Toolbox Architecture for All-IP Networks: An Ambient Networks Approach," *IEEE Wireless Commun.*, vol. 15, no. 2, Apr. 2008, pp. 8–16.

[12] W. Eddy, "At What Layer Does Mobility Belong?" *IEEE Commun. Mag.*, Oct. 2004, pp. 155–59.

[13] S. Mohanty and I. F. Akyildiz, "Performance Analysis of Handoff Techniques Based on Mobile IP, TCP-migrate, and SIP," *IEEE Trans. Mobile Computing*, vol. 6, no. 7, July 2007, pp. 731–47.

[14] A. Yegin *et al.*, "IP Mobility Orchestrator," IETF Internet draft, draft-yegin-ip-mobility-orchestrator-00, July 2014.

[15] White Paper, "Global Internet Phenomena Report," Sandvine Inc., rev. 2013-11-08, Nov. 2013.

Biographies

Alper Yegin received his B.S. in computer engineering and M.Sc. in computer science degrees from Bogazici University, Turkey, and the University of Illinois, Urbana-Champaign, in 1994 and 1997, respectively. Since 2004, he has been working for Samsung Electronics, where he is a systems architect in the Advanced Communications Lab of the DMC R&D Center. He worked in the Solaris Software Group of Sun Microsystems between 1997 and 2001, and later at NTT DoCoMo USA Labs between 2001 and 2004. His research areas include 5G mobility management and network security.

Jungshin Park is a principal engineer at DMC R&D Center of Samsung Electronics, working on 5G network architecture and mobility management protocols. He received his B.S. and Ph.D. degrees in electronics engineering from Yonsei University in 1995 and 2004, respectively. His research interests include next-generation cellular networks, low-latency wireless networks, and multi-RAT access.

Kisuk Kweon received his M.Sc. and Ph.D. in computer science degrees from KAIST, South Korea, in 2004 and 2010, respectively. Since 2010, he has been working with the DMC R&D Center of Samsung Electronics as a senior engineer. His research interests include 5G network architecture, mobility management, and the Internet of Things.

Jinsung Lee received his B.S. and Ph.D. degrees in electrical engineering from KAIST in 2003 and 2012, respectively. Since 2012, he has been with the DMC R&D Center of Samsung Electronics. His research interests include network architecture and protocol design for next-generation cellular systems and cross-layer optimization of multihop wireless networks.

IEEE COMSOC WEBINAR

18 November 2014
#IEEEComSoc

1:00 p.m. Eastern, 10:00 a.m. Pacific, 6:00 p.m. UK, 7:00 p.m. Western Europe, 6:00 p.m. GMT
REGISTER AT COMSOC.ORG/WEBINARS // COMPLIMENTARY REGISTRATION

The Role of **Spectrum Sharing in Future Wirless Networks**

Moderator: **Dexter Johnson**

Speakers:
Cristian Gomez, Spectrum Regulation and Policy Officer, Radiocommunication Bureau, International Telecommunication Union
Shih Mo, Chief Executive Officer, AviaComm

With wireless networks facing ever-increasing demand through the proliferation of smart phones, industry analysts believe that spectrum sharing will be an inevitable part of the spectrum policy of governments around the world going forward. This will require mobile operators, infrastructure suppliers and device makers to invest in the technologies and develop new business models to make spectrum sharing a reality.

Issues to be discussed:
Database Management of Spectrum
Small-cell Technology
Cognitive Radios
Regulations To Support Spectrum Sharing

Attendees will learn the technologies that will need to be employed to make this work, how government regulations are shaping up as it begins to roll out and the impact this will have for affected industries.

supported by

Anritsu

GL Communications

ROHDE&SCHWARZ

Toward D2D-Enhanced Heterogeneous Networks

Francesco Malandrino, Claudio Casetti, and Carla-Fabiana Chiasserini

ABSTRACT

In this paper, we examine upcoming 5G networks where the support of D2D communication is expected to be a key asset for operators and users alike. First, we argue the need to functionally integrate D2D and infrastructure-to-device (I2D) modes. Next, we address practical issues such as integrated resource scheduling of D2D communication within heterogeneous networks, proposing an extension of the proportional fairness algorithm that we call multi-modal proportional fairness (MMPF). We evaluate the impact of D2D in a two-tier scenario combining macro- and micro-coverage, finding that although I2D retains a clear edge for general-purpose downloading, D2D is an appealing solution for localized transfers as well as for viral content.

INTRODUCTION

These are exciting times for cellular networks. Fourth generation (4G) networks are now being deployed and offered to customers in several countries all over the world, and 5G ones are brewing. 5G is not only an evolution of current network generations, but, more significantly, a revolution in the ICT field: it will efficiently enable new ultra reliable, dependable, secure, privacy-preserving, and delay-critical services to everyone and everything, such as cognitive objects and cyber physical systems.

Among the new features foreseen by 5G networks, device-to-device (D2D) transfers may have a prominent role, and our purpose here is to assess whether, and to what extent, they can be successfully integrated into 5G networks.

The adoption of D2D transfers is driven by four main use cases [1]:

- Safety applications and disaster scenarios
- Novel commercial proximity services (ProSe) scenarios
- Network traffic offloading
- Industrial automation and machine-to-machine communication [2]

The focus of our work is on the use of D2D for commercial applications, and, in particular, for data traffic offloading. The typical example is "flash crowds" [3]: thousands of users in a small area, perhaps attending a football game, suddenly becoming interested in the same content (e.g., a replay clip). Some users may download the content through D2D, thus partly relieving the infrastructure of its load. Furthermore, through reduced interference and extended coverage, D2D may lead to increased network capacity.

Cellular networks are not the first technology to allow D2D transfers: similar techniques, aiming at roughly the same goals, such as ad hoc mode in IEEE 802.11, have existed for decades, but have never become mainstream. Technical and non-technical issues, from driver support to security concerns, have always hindered their widespread adoption.

The first important question we have to ask ourselves then is: why should we expect D2D to be successfully ushered into cellular networks? For decades, they have been working — quite successfully, in fact — in an infrastructure-centric fashion: users send data to base stations, base stations send data to users. Should we dare depart from such a reliable, tested working scheme in exchange for some performance improvement and dire technical challenges?

The second, related question: do not small-cell techniques serve the same goals as D2D, i.e., reduced interference, extended coverage, and network offloading? They certainly allow operators to retain the familiar infrastructure-centric operation mode. Does this mean that we can just use small cells and disregard D2D? Are there use cases and scenarios where D2D is more appropriate? More interestingly, do we need to choose between small cells and D2D, or can the two paradigms coexist? And if they do, at what cost in terms of complexity and overhead?

It is important to stress that we are not investigating here the usefulness of D2D itself, but rather the opportunity for its integration in cellular networks. Although D2D is now part of 3GPP standards, this does not necessarily imply that it will be implemented by vendors and operators, nor that it will be widely used in services and applications.

We explore these fundamental questions with a focus on the integration of D2D in cellular networks. Then we introduce a system architecture for D2D support. We describe the problem

The authors are with Politecnico di Torino.

0163-6804/14/$25.00 © 2014 IEEE

of radio resource sharing in systems where both the D2D and infrastructure-to-device (I2D) paradigms (the latter including small cells) are implemented. We present our reference scenario and our numeric results. Finally, we draw our conclusions.

D2D in Heterogeneous Cellular Networks

In this section, we discuss two important preliminary issues. First, we analyze how D2D communication should be integrated in cellular networks. Then we discuss whether D2D and small cell technologies are alternative or complementary to each other.

Integrating D2D in Cellular Networks

As discussed in [4, 5], there are three main options concerning the integration of D2D in cellular networks:

- Whether D2D communication shall be network-controlled or not
- Whether it shall happen in-band or out-of-band
- Whether it shall work in overlay or underlay fashion

For a conceptual framework concerning problems such as peer discovery, scheduling, and resource allocation, see [6].

In the following, we discuss the most promising solutions, assuming that D2D-enhanced cellular networks will be network-controlled and operate in-band in an underlay fashion. As presented in [2], this is a fairly popular choice.

Network-controlled D2D essentially means that infrastructure nodes (base stations and control entities, as detailed later) play a central role in establishing, arbitrating, and managing D2D connections [7]. As explained in [1], they will provide the following fundamental services: spectrum management, security, information brokering, and mobility management. The first two items in this list guarantee that D2D transfers do not translate into lower performance or poorer security. The last two imply that user equipment (UE) does not even have the burden of choosing between D2D and I2D. The opposite approach is represented by infrastructure-less networks, such as WiFi Direct or Bluetooth.

In-band D2D refers to the fact that D2D traffic uses the same licensed frequencies as ordinary I2D traffic [7]. The main advantage of this approach is represented by the higher degree of control operators retain on who transmits and how, which limits interference. Cooperation among users is also easier to enforce and check. Furthermore, terminals do not need to carry additional radio interfaces. The opposite approach, out-of-band D2D, is envisioned by those proposals that seek to offload cellular networks through other networks, such as 802.11 domestic access points with spare capacity.

Underlay refers to the fact that D2D communication has no part of the spectrum specifically reserved to it, as happens, for example, in [5]. When networks operate in underlay fashion, D2D transfers share the same radio resources as those used by traditional cellular communications and are scheduled within the cellular bands in an opportunistic fashion. As a consequence of network-controlled in-band operation, D2D transfers are *scheduled*, as mentioned earlier: this avoids the potential inefficiency of decentralized schemes based on the carrier sense multiple access (CSMA) paradigm. The overlay approach, instead, implies that separate radio resources are devoted to D2D and I2D, or to device-to-infrastructure (D2I) communication.

D2D and Small Cells

Small cell is an umbrella term, covering several, quite different, technologies. In general, it refers to low-power short-range operator-owned nodes integrated in the cellular infrastructure to enhance its coverage and/or capacity. Such communication nodes go under names like pico- or micro-eNBs. The goals of small cells are very similar to those of D2D, so one may think we just need to choose the most effective among the two and discard the other.

There is, however, a very important difference, and it concerns the *source* of the information being transmitted. With small cells, information is still downloaded from some remote server (i.e., on the Internet) and then transmitted to the user. With D2D, instead, the information moves directly from one user to another, either generated by the transmitting user itself, or previously received from a remote server or another user. With reference to the use cases described earlier, we can say that both D2D and small cells can be used for network offloading, but D2D is the most effective choice for ProSe scenarios. Later in this work, we study how the two can coexist in a real-world network.

System Overview and Architecture

Due to the new complex tasks assigned to the cellular infrastructure (arbitrating and scheduling D2D transfers), we envision that base stations (eNBs in LTE terminology) will be assisted by a new kind of entity: area controllers (ACs). The main difference between these entities is the level at which they operate. eNBs are solely concerned with propagation and spectrum aspects. Area controllers, instead, have a wider view of the network: they control a bigger area, and are in charge of content-aware decisions. More exactly:

- eNBs assign individual spectrum resources (PRBs in LTE) to pairs of communicating endpoints (i.e., themselves and a user, or a pair of users through D2D).
- ACs monitor the network state, including content demand and propagation conditions, and exploit such information to decide the paradigm (I2D or D2D) and the amount of resources a transfer should be assigned.

The system model, along with the way ACs, eNBs, and users interact with each other, is summarized in Fig. 1.

Due to the new complex tasks assigned to the cellular infrastructure (arbitrating and scheduling D2D transfers), we envision that base stations (eNBs in LTE terminology) will be assisted by a new kind of entity: area controllers.

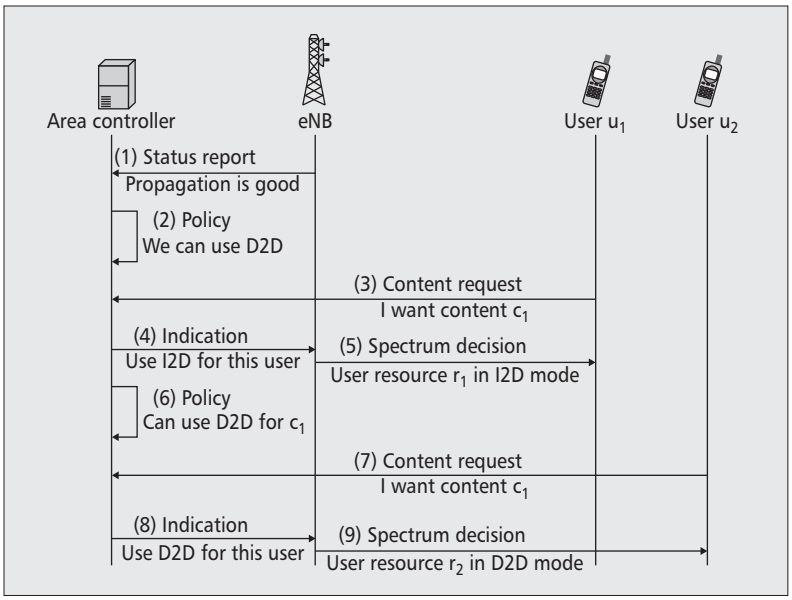


Figure 1. System model, summarizing the entities involved in heterogeneous networks, and interaction diagram highlighting the role of each entity.

eNBs update the AC on propagation conditions (1). The AC uses such information to update its policies (2). Upon a request from users (3), the AC is also in charge of making content-aware decisions (4), such as whether certain content should be downloaded through D2D or I2D. Although eNBs do have better knowledge of propagation conditions than ACs, they are not aware of which content items are stored by each user, and thus which D2D transfers are possible. eNBs are, however, in charge of actual scheduling decisions (5).

Using its knowledge of the content being downloaded in the network, the AC can refine (6) its policies (e.g., by understanding that now content c_1 can be downloaded from user u_1 using D2D if needed). Such decisions are transmitted (8) to the eNB, which subsequently enacts them (9) using the most appropriate spectrum resources.

Operations performed by eNBs have, above all, to be fast: in LTE, PRBs are assigned as frequently as once per millisecond. Furthermore, being part of the access network, eNBs are oblivious to higher-level concerns: as an example, they cannot run content-aware scheduling algorithms. Such algorithms do, however, accept input parameters, such as which users should get their content through D2D. Conversely, ACs are in an ideal position to make more complex decisions, exploiting more information and (possibly) taking longer to find the (ideally) optimal one. The interface between ACs and eNBs consists in the parameters of the fine-grained scheduling algorithms run by eNBs.

It is worth stressing that the timescales at which ACs and BSs operate are different. More exactly, while eNBs have to make decisions each millisecond, ACs can refine their policies over longer time intervals, accounting for more information and making more accurate (and, if need be, computationally complex) decisions.

Integrated I2D and D2D Scheduling

This section is chiefly concerned with how D2D in LTE will happen. Specifically, we address spectrum and bandwidth aspects and scheduling.

Spectrum, Bands and Resources

Unlike previous-generation networks such as 3G, LTE and its successors operate on several different carrier bands, anywhere between 700 MHz and 3.8 GHz in frequency, and anywhere between 1.4 and 20 MHz in width. This allows flexible scalable deployments, with wider lower-frequency cells coexisting with smaller higher-frequency ones. The availability of specific bands changes on a regional and country basis.

Details of the frequency and time structure of LTE are fairly complex, and are beyond the scope of this article. Here we simply refer to physical resource blocks (PRBs). Physical resource blocks correspond to a set of resources in the time and frequency domains, that is, a bandwidth of 180 kHz for the duration of a time slot (0.5 ms), which corresponds to half a subframe. The subframe is the atomic scheduling unit for LTE networks. As an example, let us consider a bandwidth of 10 MHz for the downlink, which accommodates 50 PRBs per time slot available for data traffic.[1] In downlink each PRB can carry up to 504 bits; then the actual data rate depends on propagation conditions, interference, and on the use of multiple-input multiple-output (MIMO). From our viewpoint, scheduling simply means deciding how many PRBs shall be assigned to each pair of endpoints (an eNB and a UE, or two UEs) that need to communicate.

Given the uplink and downlink subcarriers, one may wonder where D2D would fit in such a picture. Indeed, the topic has been widely debated in standardization fora; the predominant orientation is to accommodate D2D traffic in uplink resources [4]. Reasons for this choice include lower interference and reduced impact on eNB-to-UE links. Additionally, the uplink spectrum is currently underutilized [4]. Thus, it is more appropriate to use it for content downloading use cases (including flash crowds), where D2D links can effectively be used for network offloading.

Scheduling in Heterogeneous Networks

Scheduling in next-generation heterogeneous networks is a complex task. Essentially, we have to assign a finite set of resources (PRBs) to a set of pairs of endpoints. These pairs of endpoints (i) may want to communicate in I2D or D2D fashion; (ii) differ in position, mobility, and propagation conditions; (iii) aim at fetching different amounts of different content. The very metric to optimize is unclear: one may think that the total network throughput is a good candidate; however, this would mean disregarding fundamental fairness issues.

The scheduling algorithm used in virtually all current networks is proportional fairness (PF). Users are given a priority that is directly proportional to the rate they can achieve and inversely

[1] The remaining bandwidth is used as guard bands.

proportional to the amount of data they have already transferred. The overall effect is to allow users with better propagation conditions to transmit more data (thus enhancing the global network capacity), without starving the others (thus guaranteeing a certain level of fairness).

Traditional PF is only concerned with *which* users shall be served, but not with how to do so. Indeed, in traditional cellular networks there is only one way to serve users: through the closest base station. In heterogeneous networks, however, we have two different problems:

- User service mode (i.e., through macro- or micro-eNBs, or via D2D)
- Resource allocation

Not surprisingly, dealing with these two issues at the same time is substantially more complex than traditional scheduling. Many D2D-aware scheduling algorithms have been proposed for LTE networks [2, 8, 9], with different strategies and objectives. However, in this article we are chiefly concerned with designing a practical scheduling scheme that can be implemented in cellular networks and with assessing the sheer impact of D2D on the performance of cellular networks. Thus, instead of presenting an optimized scheduling algorithm accounting for all the issues (and opportunities) of heterogeneous, D2D-enhanced networks, we propose an evolution of the PF algorithm and study its performance with and without D2D support. We name this algorithm Multi-Modal Proportional Fairness (MMPF).

The MMPF Algorithm

As mentioned, scheduling in heterogeneous networks entails making two decisions: which users we should serve, and how to do it. The first decision is the same as in traditional proportional fairness, and therefore is made in the same way. At each step, we select the user with the highest ratio between achievable rate and amount of already downloaded data. As for the second decision, we proceed as follows:

- If there is another user with the requested content within a distance R_{max}, use D2D.
- Otherwise, use I2D.

The pseudocode of the MMPF algorithm is presented in Fig. 2.

Note that if I2D is selected, either macro- or micro-eNBs can be used, whichever has the highest received signal strength indicator (RSSI) (line 9). We may observe a lack of symmetry: we decide whether to use D2D based on distance, but then select macro- or micro-eNBs based on signal quality. The reason is eminently practical. While eNBs (and thus the AC) have many ways of estimating the location of users [10], there is no simple way they can obtain reliable information on the signal quality between two users. Estimation techniques do exist, but none is included in standards and all of them imply some overhead.[2] Similarly, using the threshold R_{max} is a somewhat coarse, hardly optimal, way of choosing between D2D and I2D. Other, more sophisticated approaches may yield a higher average throughput, better fairness, or both. As an example, we may reuse the resources allocated for D2D, due to the lower interference they suffer. Recall, however, that our purpose is to assess the impact of D2D on the performance of cellular networks, and to which extent it can replace — or complement — small cell solutions. In view of this, having an algorithm that closely resembles the de facto standard of today's network is particularly appropriate and convenient.

```
1: for all user u do
2:     compute u's score, given by
                    u's rate
       ------------------------------------
       data downloaded by u
3: for all RB r do
4:     let u* be the user with the highest score
5:     if the content needed by u* is available through d2d then
6:         let s* be the closest user to u* having the content she needs
7:         schedule RB r for transfer s* → u* through d2d
8:     else
9:         let b* be macro- or micro-BS covering u* with the best RSSI
10:        schedule RB r for transfer b* → u* through i2d
```

Figure 2. The MMPF algorithm.

It is important to point out that the first decision made by MMPF, whether to use I2D or D2D, also impacts the part of spectrum (uplink or downlink) over which the transfer occurs. The original PF scheduling is applied independently for uplink and downlink; it follows that, in the uplink bandwidth, D2D downloads and D2I uploads will be scheduled together (i.e., will compete with each other). Therefore, the amount of upload traffic will have an impact on the performance of D2D downloads and vice versa.

Unlike the original PF algorithm, MMPF is run at the AC, as described earlier. Indeed, the second decision whether to use D2D or I2D, has to be content-aware and must be made by an entity with more complete information about the content available at each user. It is also important to stress that users are never requested to download a content item in which they are not interested for the sole purpose of acting as relays. Users are requested to share the content they have already downloaded but nothing more.

It is also worth pointing out that, as the name suggests, MMPF is fair in the same sense the original proportional fairness algorithm is. As we can see in line 2, the next user to serve is selected taking into account the ratio between the rate of each user and the amount of data he/she already downloaded. As in the original algorithm, this does not imply that all users end up receiving the same amount of data — another, more intuitive, definition of fairness.

Reference Scenario

We evaluate our solution in the two-tier scenario typically used within 3GPP for LTE network evaluation [11]. The scenario comprises a service network area of 12.34 km^2, covered by 57 macrocells and, unless otherwise specified, 228 microcells. Macrocells are controlled by 19 three-sector macro-eNBs; the macro-eNBs intersite distance is set to 500 m. Micro-eNBs are deployed over the network area so that there are

[2] Other simulations, when the quality of D2D links is perfectly known and used in MMPF, yielded essentially the same results as the ones we present here.

Content class	Number of content items	Size (Mb)	Deadline (s)	Request rate (copies/ms)
Software updates	10	12	4	0.1
Video	10	3	1	0.1
Viral	1	3	1	50

Table 1. Content classes.

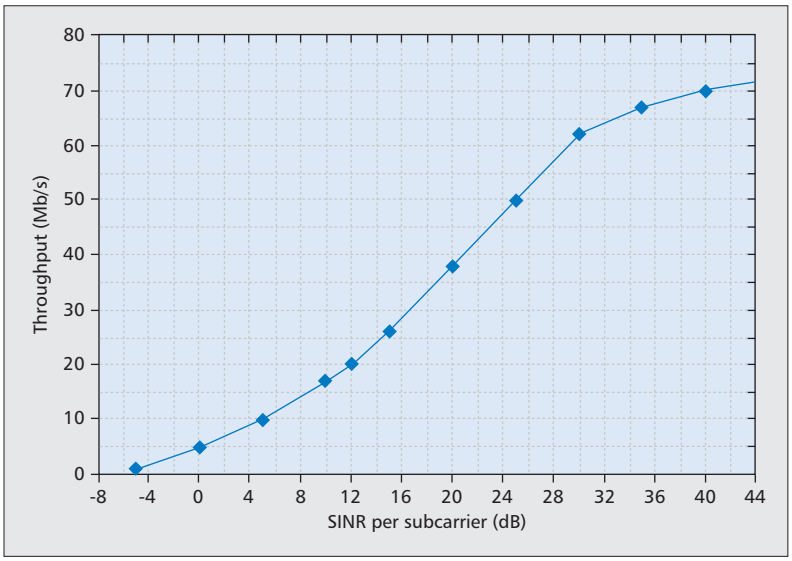


Figure 3. Throughput vs. SINR, based on the experimental results in [14].

four non-overlapping microcells per macrocell. A total of 3420 users are present in the area. In particular, in order to have a higher user density where microcells are deployed, 10 users are uniformly distributed within 50 m from each micro-eNB. The rest of the users are uniformly distributed over the remaining network area. Users move according to the cave man model [12], with average speed of 1 m/s.

According to current specifications [11,13], we assume the following pairs of values for power and antenna height: (43 dBm, 25 m) for macro-eNBs, (30 dBm, 10 m) for micro-eNBs, and (23 dBm, 1.5 m) for UEs. All network nodes operate over a 10 MHz band at 2.6 GHz; thus, there are a total of 50 PRBs to assign for each subframe. As already mentioned, the signal propagation is realistically modeled according to International Telecommunication Union (ITU) specifications for urban environment [11], while the signal-to-interference-plus-noise (SINR) is mapped onto peak throughput values using the experimental measurements in [14]. More precisely, the propagation loss for macrocells, microcells, and D2D is given by the following expressions [11]:

- $13.5 + 20\log_{10}f_c + 39\log_{10}d$
- $22.7 + 26\log_{10}f_c + 36.7\log_{10}d$
- $27.0 + 20\log_{10}f_c + 22.7\log_{10}d$

where f_c is the carrier frequency, and d is the distance between transmitter and receiver. The experimental measurements of [14] are summarized in Fig. 3, and refer to the case of 2 × 2 MIMO.

	Macro	Micro	D2D
PF	0.96	0.81	
MMPF	1.29	1.18	0.6

Table 2. Spectral efficiency for different transfer paradigms.

Users choose the content they request from a set of 21 different items belonging to three categories: software updates, videos, and viral content [15]; their size, deadline, and request rates are summarized in Table 1. Content items belonging to each category are chosen with uniform probability; similarly, the request rate is uniform throughout the simulation. Content deadlines imply that service requests are aborted after the deadline expires. Notice, however, that MMPF, just like original PF, does not account for such deadlines while scheduling traffic. We highlight that video and viral items have stricter constraints on delivery time. Additionally, the viral item is modeled as being in high demand to mimic content becoming suddenly popular through social networks because of "flash crowds." We assume that a scheme based on incentives is implemented, and thus that users are willing to cooperate by providing content upon receiving a request.

Simulations are carried out through a custom simulator. The total simulation time is 30 s.

It is important to note that the scenario we are addressing will necessarily only paint a part of the overall picture. Additional case studies involving out-of-coverage communication as well as interactive and conversational services should be addressed in future work for reasons of space.

Results

We vary the maximum distance allowed for D2D transfers, R_{max}, between 10 and 100 m, and study the amount of data that is transferred through each of the possible paradigms: macrocells, microcells, and D2D. We are interested in investigating how D2D relates to the other paradigms, especially microcells. The results are depicted in Fig. 4. First of all, we observe that most of the data flow through microcells; this is expected as this paradigm offers an excellent balance between short-range (i.e., low interference) and high power. A smaller amount of data flows through macrocells, and a more limited amount through D2D.

As the maximum range allowed for D2D increases, so does the amount of data transferred through it. This is essentially because a wider range directly translates into a higher number of potential sources (i.e., users carrying the needed content). More interestingly, increasing the range also increases the amount of data transferred through the other paradigms. Indeed, allowing more room for D2D has two positive effects:

- Some data can be downloaded from closer sources, hence with higher quality and lower interference.

• D2D transfers occur on uplink bands; thus downlink I2D traffic, hence interference, is reduced. Both imply that more downloads are completed within their deadline, which explains why the overall amount of transferred data increases.

In other words, by allowing D2D, not only can we move some traffic to the uplink spectrum, but we can also more effectively use the downlink one, increasing its capacity. Congestion is, of course, increased on the uplink bands. However, with most of the present (and future) cellular traffic being represented by downloads, this is an acceptable trade-off.

Table 2 summarizes the spectral efficiency, that is, the amount of data transferred through each PRB, for each traffic paradigm. We clearly see that by allowing D2D transfers we substantially increase the efficiency of macrocell- and microcell-based I2D as well.

Next, we try to assess whether D2D can complement, or replace, microcells. To this end, we reduce the number of microcells by 50 percent: this is not a disaster scenario, but it may account for, say, a less pervasive deployment due to economic concerns. The effect is summarized in Fig. 5. The first obvious observation is that there is much less traffic flowing through microcells. Such a loss is compensated partly by macrocells and partly by D2D, although the overall network capacity decreases. Thus, it cannot be said that D2D can fully replace microcells (and spare the costs of deploying that kind of infrastructure); nevertheless, D2D can complement microcells and offset a significant part of the effects of a reduced infrastructure deployment.

The above observation prompts another question: are there some content items that are more suitable than others for D2D delivery? The answer can be obtained by looking at which content gets transferred through each paradigm for the baseline scenario with a 50-m limit distance for D2D. From Fig. 6, it can be observed that D2D is especially effective at transferring viral content. Indeed, viral content is highly popular, and all requests for it happen in a fairly short interval of time; thus, it is much more likely to find a neighbor device that can provide the requested content.

In summary, we can say that D2D is an extremely appealing solution for proximity services and offloading of highly popular content. Although it cannot fully replace the ordinary I2D paradigm (including microcells) for general-purpose downloading, it can be successfully integrated within heterogeneous cellular-based networks. To answer the question we raised earlier about the coexistence between D2D and small cell approaches, we can conclude that these two paradigms serve different, albeit partially overlapping, purposes, and can — and should — profitably be integrated together in future cellular networks.

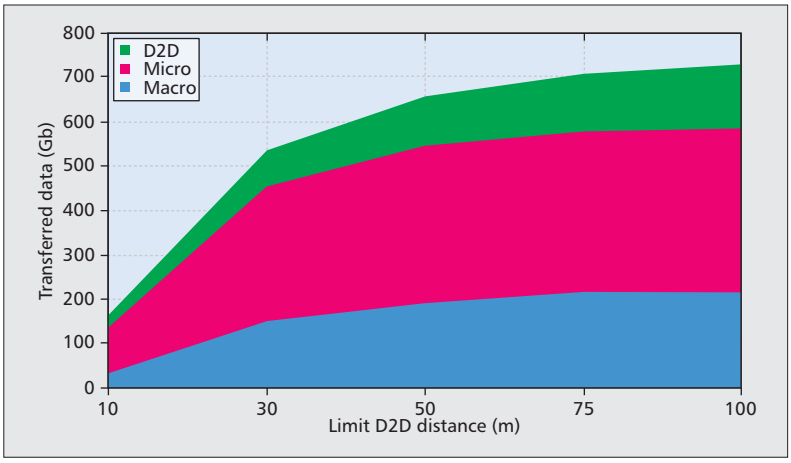


Figure 4. Baseline scenario: amount of transferred data through each paradigm.

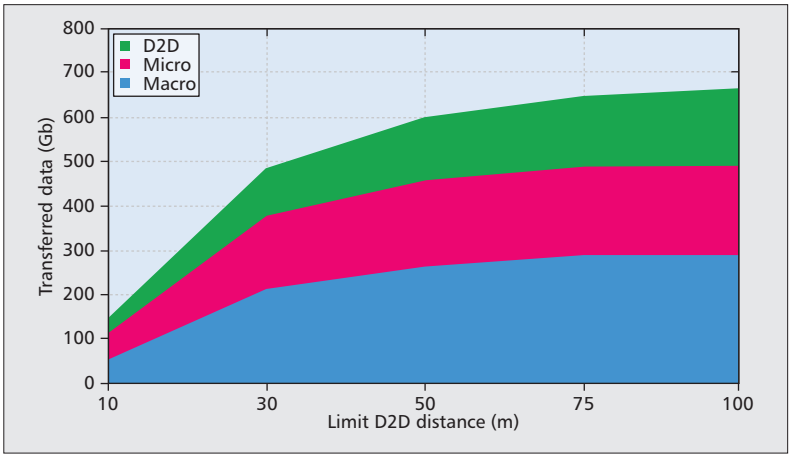


Figure 5. Amount of transferred data for each paradigm when the number of microcells is halved.

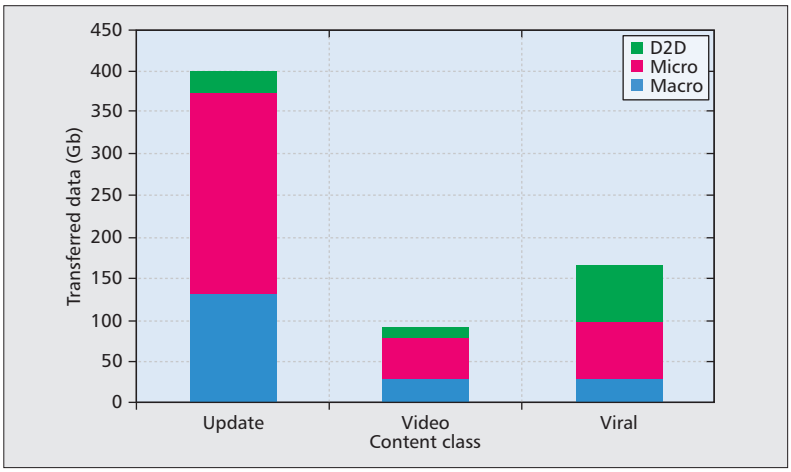


Figure 6. Data transferred through each paradigm for different content classes.

Conclusions

The main contribution of our work is the analysis of integrated D2D and I2D scenarios. We discuss motivations for the coexistence and joint deployment of both transfer modes, as well as a practical solution for resource scheduling in a multimode context that extends the popular proportional fairness scheduling algorithm. Performance evaluation, in a typical scenario used for LTE network studies, provides new grounds for incorporating D2D in future releases of 3GPP

D2D is an extremely appealing solution for proximity services and offloading of highly popular content. Although it cannot fully replace the ordinary I2D paradigm (including microcells) for general-purpose downloading, it can be successfully integrated within heterogeneous cellular-based networks.

standards by showing its effective offloading potential in case of both localized and highly popular content.

With reference to the questions posed in the Introduction, it can be concluded that:

- D2D can indeed be profitably integrated within cellular networks.
- It cannot fully replace small cells but is useful to complement them and mitigate the effects of a reduced infrastructure deployment.

Acknowledgment

This article was made possible by NPRP grant #5-782-2-322 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] C. Mouton, "Device-to-Device Standardization in 3GPP," RAS Cluster meeting, Lisbon, Portugal, July 2013.
[2] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," to appear, *IEEE Commun. Surveys and Tutorials*.
[3] T. Broxton *et al.*, "Catching a Viral Video," *IEEE ICDM Wksps.*, 2010.
[4] X. Lin *et al.*, "An Overview on 3GPP Device-to-Device Proximity Services," ArXiV preprint 1310.0116.
[5] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum Sharing for Device-to-Device Communication in Cellular Networks," ArXiV preprint 1305.4219.
[6] G. Fodor, E. Dahlman, and G. Mildh, "Design Aspects of Network Assisted Device-to-Device Communications," *IEEE Commun. Mag.*, 2012.
[7] L. Lei *et al.*, "Operator Controlled Device-to-device Communications in LTE- Advanced Networks," *IEEE Wireless Commun.*, 2012.
[8] A. Asadi and V. Mancuso, "On the Compound Impact of Opportunistic Scheduling and D2D Communications in Cellular Networks," *ACM MSWiM*, Barcelona, Spain, Nov. 2013.
[9] F. Malandrino *et al.*, "Fast Resource Scheduling in HetNets with D2D Support," *IEEE INFOCOM*, Toronto, Canada, May 2014.
[10] S. Cheiran and A. Rudrapatna, "A Primer on Location Technologies in LTE Networks," *Alcatel-Lucent Techzine*, http://www2.alcatel-lucent.com/techzine/a-primer-on-location-technologies-in-lte- networks.
[11] 3GPP Tech. Rep. 36.814, "Further Advancements for E-UTRA Physical Layer Aspects," 2010.
[12] D. J. Watts, "Small Worlds: The Dynamics of Networks between Order and Randomness," *Princeton Studies on Complexity*, 1999.
[13] ITU-R, "Guidelines for Evaluation of Radio Interface Technologies for IMT-Advanced," Rep. ITU-R M.2135-1, Dec. 2009.
[14] D. Martin-Sacristan *et al.*, "3GPP Long Term Evolution: Paving the Way towards Next 4G," *Waves*, 2009.
[15] T. Lohmar *et al.*, "Delivering Content with LTE Broadcast," *Ericsson Rev.*, 2013.

Biographies

Francesco Malandrino graduated (summa cum laude) in computer engineering from Politecnico di Torino in 2008. Since 2009, he has been with the Telecommunication Networks Group at Dipartimento di Elettronica of Politecnico di Torino, first as a Ph.D. student, then (sice 2012) as a post-doctoral fellow. From 2010 till 2011, he was a visiting researcher at the University of California at Irvine. His interests focus on wireless and vehicular networks and infrastructure management.

Claudio Casetti [M'05] graduated from Politecnico di Torino in 1992 and received his PhD in electronic engineering from the same institution in 1997. He is an assistant professor at Politecnico di Torino. He has coauthored more than 130 papers in the field of networking and holds three patents.

Carla-Fabiana Chiasserini [M'98, SM'09] received her Ph.D. in 2000 from Politecnico di Torino, where she is currently an associate professor. Her research interests include protocols and performance analysis of wireless networks. She has published over 200 papers at major venues, and serves as Associate Editor of several journals.

LEVERAGING TECHNOLOGY - THE JOINT IMPERATIVE

MILCOM 2015

CALL FOR PAPERS

OCTOBER 26-28, 2015 • TAMPA FLORIDA, USA

AFCEA

IEEE COMMUNICATIONS SOCIETY

IMPORTANT DATES

TECHNICAL PAPERS

DRAFT PAPERS DUE:
April 6, 2015

PAPER ACCEPTANCE NOTIFICATION:
June 4, 2015

FINAL PAPERS DUE:
July 20, 2015

PRESENTATIONS DUE:
August 15, 2015

TUTORIALS AND PANELS

PROPOSALS DUE:
April 6, 2015

ACCEPTANCE NOTIFICATION:
June 4, 2015

MATERIALS DUE:
August 15, 2015

MILCOM 2015 solicits unclassified technical papers and proposals for tutorials and panels on current and emerging topics applicable to all facets of military communications. We encourage professionals in industry, academia and government worldwide to contribute and participate. The general forum will be open to all and there will be a restricted access portion of the program to support ITAR-controlled or FOUO material.

TECHNICAL TRACKS - Topics include but are not limited to the following:

TRACK 1 – WAVEFORMS AND SIGNAL PROCESSING

- Advanced Antenna and RF Technology
- Anti-jamming Techniques
- Cognitive Radios
- Compressive Sensing
- Cooperative Communication
- Dynamic Spectrum Management
- Free-space Optical and Laser-based Communications
- Modified Commercial Wireless Communication
- Modulation and Coding
- Physical-layer Security
- Propagation and Channel Modeling
- Radar System, Detection and Localization
- Satellite Signals
- Signal Processing Algorithms
- Underwater Communications

TRACK 2 – NETWORKING PROTOCOLS AND PERFORMANCE

- Ad Hoc, Mesh, Sensor Networks
- Capacity Analysis and Optimization
- Cognitive Networks
- Content-Based Networking
- Cross-Layer Design
- Disruption-Tolerant Networks
- Heterogeneous Networks
- Hybrid Optical/RF Networks
- Multicasting
- Network and Information Sciences
- Network Coding
- Network Discovery
- Networked Control System
- Optical Networks
- Overlay Networks
- Protocols: MAC, Link-Layer, Network, Transport
- Quality of Service
- Routing
- Software Defined Networking
- Topology and Network Control

TRACK 3 – CYBER SECURITY AND TRUSTED COMPUTING

- Authentication, Authorization and Accounting
- Access Control
- Cloud Security
- Critical Infrastructure Security
- Cyber Operations
- Cyber Defense
- Distributed Systems Security
- Identity Management
- Intrusion Detection
- Joint Tactical Network Security
- Malware Analysis
- Mobile Network Security
- Risk Management
- Threat Monitoring and Analysis
- Trust and Privacy

TRACK 4 – SYSTEM PERSPECTIVES

- A2/AD Systems and Operations
- Airborne Networks
- Autonomous Systems
- Communications-on-the-Move
- Disadvantaged Networks
- IED Detection/Geolocation
- Integrated EW and Communications
- Satellite Communications and Networks
- Small Satellite Communications Networks
- System Modeling and Simulations
- Space Networks
- System Architecture
- Test Beds, Experiments, Exercises and Demonstrations

TRACK 5 – SELECTED TOPICS IN COMMUNICATIONS

- Cognitive Analysis
- Communication SW, Services, and Multimedia Applications
- E-skin Communications
- Homeland Security
- Human Behavior in Cyber World
- Human Factors in Communication System Design
- Internet of Things
- Low-Power Devices
- LPI/LPD/LPE Communications
- Machine-to-Machine Communications
- Nano Communication Networks
- Network Science
- Sensor Communications
- Social Networks
- Survivable Communications and Networks

TECHNICAL PAPERS

Unclassified Technical Program

The unclassified technical program provides a venue for papers and presentations that do not include ITAR-sensitive, classified or proprietary information.

Restricted-Access Technical Program

The restricted-access technical program provides a venue for papers that contain ITAR-controlled or FOUO information. To attend and participate in the restricted-access program, one must be a U.S. citizen. In addition, papers originated outside of U.S. will not be accepted in this program.

TUTORIALS AND PANELS

MILCOM 2015 also solicits proposals for unclassified half-day tutorials and one-to two- hour panels. Relevant topics include state-of-the-art communications and networking technologies, cyber, and innovative techniques applied to communication systems and acquisition strategies for joint imperative.

ADDITIONAL INFORMATION

Specific submission and proposal information, including information about travel grants for students, is available at www.milcom.org.

PROGRAM CHAIRS

Technical Program Chair
Qinqing (Christine) Zhang, Ph.D
University of Southern California (*qinqing@ieee.org*)

Unclassified Technical Program Chair
Jerry Brand, Ph.D, PE
Harris Corporation (*j.brand@ieee.org*)

Restricted-Access Technical Program Chair
Thomas Macdonald, Ph.D
MIT Lincoln Laboratory (*tom_macdonald@ll.mit.edu*)

Technical Program Vice Chairs
Matthew Valenti, Ph.D
West Virginia University
(*Matthew.Valenti@mail.wvu.edu*)

Harlan Russell, Ph.D
Clemson University (*hrussel@clemson.edu*)

Tutorials Program Chair
Bharat Doshi, Ph.D
Johns Hopkins University, Applied Physics Lab
(*bharat.doshi@jhuapl.edu*)

Panels Program Chair
Bonnie L. Gorsic
The Boeing Company (*bonnie.l.gorsic@boeing.com*)

GREEN COMMUNICATIONS AND COMPUTING NETWORKS




Jinsong Wu *John Thompson* *Honggang Zhang* *Daniel C. Kilper*

Information and communication technologies (ICT) widely contribute to the global economy and society through the tremendous pace of innovations and new applications, rapidly changing the way that people live in almost every aspect. However, ICT can also have a profound impact on the global environment, which could be either positive or negative. This IEEE Series on Green Communications and Computing Networks is now established to discuss the concepts, principles, mechanisms, design, algorithms, analyses, and research challenges relevant to ICT and their environmental impacts, which could be referred to as "green ICT." Here the term green is not only used to include energy or energy efficiency issues, but is also considered in the broader context of environmental impact and enablement of sustainability through communication and computing networks. A common misunderstanding is that green is simply equivalent to energy or energy efficiency issues. Actually, the field will explore more and more green topics relevant to issues other than energy or energy efficiency. Thus, this Series will not only address green communications, green computing, and relevant systems, but also investigate using communications, computing, and relevant technologies to achieve green objectives for a sustainable world.

This, the inaugural issue of this Series, includes four articles that mostly discuss energy-relevant green issues, although as indicated above, this Series also solicits and encourages contributions on non-energy-related green topics.

The invited article "Simultaneous Wireless Information and Power Transfer in Modern Communication Systems," written by I. Krikidis *et al.*, provides an overview of simultaneous wireless information and power transfer (SWIPT) systems with a particular focus on the hardware realization of rectenna circuits and practical techniques. This article also discusses the benefits from a potential integration of SWIPT technologies for resource allocation and cooperative cognitive radio networks.

The article "Green Transmission Technologies for Balancing the Energy Efficiency and Spectrum Efficiency Trade-off," written by Y. Wu *et al.*, provides four selected green transmission technology (GTT) solutions, focusing on how they utilize the degrees of freedom in different resource domains, as well as how they balance the fundamental trade-off between energy efficiency and spectrum efficiency. Furthermore, this article also introduces the GTT toolbox as a systematic tool and unified simulation platform for the proposed GTT solutions.

The third article, "A Survey of Energy-Efficient Caching in Information-Centric Networking" by C. Fang *et al.*, offers a brief survey of energy-efficient caching techniques in information-centric networking (ICN) from cache placement, content placement, and request-to-cache routing perspectives. This article also discusses some relevant challenges and future research directions about caching policies for green ICNs.

Finally, the article "Approaches to Energy Intensity of the Internet," written by D. Schien and C. Preist, considers the approaches of top-down and bottom-up modeling to estimate the network energy intensity in theInternet, and reviews the varying assumptions in existing bottom-up models and combines them in a meta-mode, which might provide more robust estimates of the approximate energy efficiency for networks.

ACKNOWLEDGMENTS

We would like to acknowledge the great support and help from Sean Moore, the Editor-in-Chief of *IEEE Communications Magazine*, Charis Scoggins, Administrative Aide to the Editor-in-Chief, Jennifer Porcello, Production Specialist, and the other IEEE Communications Society publication staff. We also highlight the great support for this Green Series from the members of the Technical Committee on Green Communications and Computing (TCGCC) of the IEEE Communications Society.

BIOGRAPHIES

JINGSON WU [SM] (wujs@ieee.org) is the founder and Founding Chair of the Technical Committee on Green Communications and Computing (TCGCC), IEEE Communications Society (established as Technical Subcommittee on Green Communications and Computing, TSCGCC, in 2011, elevated to TCGCC in 2013). He is an Associate Editor of *IEEE Communications Surveys & Tutorials*, *IEEE Systems Journal*, and *IEEE Access*, and a Series Editor of

IEEE Series on Green Communications and Computing Networks for *IEEE Communications Magazine*. He has been a Guest Editor of the *IEEE Systems Journal* Special Issue on Green Communications, Computing, and Systems, IEEE Access Special Section on Big Data for Green Communications and Computing, and *Elsevier Computer Networks Journal* Special Issue on Green Communications. He was lead General Chair of the IEEE International Conference on Green Computing and Communications 2013, and Technical Program Committee Co-Chair of the IEEE Online Conference on Green Communications in both 2012 and 2013. He was the leading Editor and co-author of the comprehensive book *Green Communications: Theoretical Fundamentals, Algorithms, and Applications* (CRC Press, 2012). He received his Ph.D. in electrical and computer engineering from Queen's University, Kingston, Canada.

JOHN THOMPSON [SM] (john.thompson@ed.ac.uk) currently holds a personal chair in Signal Processing and Communications at the School of Engineering in the University of Edinburgh. His research interests currently include signal processing, energy-efficient communications systems, and multihop wireless communications. He was deputy academic coordinator for the recent Mobile Virtual Centre of Excellence Green Radio project, which involved collaboration between five U.K. universities and a dozen international companies. He currently leads the European Marie Curie Training Network ADVANTAGE, which trains 13 Ph.D. students in the area of smart grid technology. During 2012–2014 he is serving as a Member-at-Large for the Board of Governors of the IEEE Communications Society. He is also a distinguished lecturer for ComSoc in 2014–2015. He was Technical Program Co-Chair for the IEEE Vehicular Technology Conference-Spring in Dresden in 2013.

HONGGANG ZHANG [SM] (honggangzhang@zju.edu.cn) is an International Chair Professor, CominLabs Excellence Center, Université Européenne de Bretagne (UEB) & Supélec, France, a full professor of the Department of Information Science and Electronic Engineering as well as co-director of the York-Zhejiang Lab for Cognitive Radio and Green Communications at Zhejiang University, China. He is an honorary visiting professor at the University of York, United Kingdom. He received his Ph.D. degree in electrical engineering from Kagoshima University, Japan, in March 1999. He was the principal contributor for proposing DS-UWB in the IEEE 802.15 WPAN Standardization Task Group. He served as the Chair of the Technical Committee on Cognitive Networks (TCCN) of the IEEE Communications Society during 2011–2012. He was Co-Chair of the IEEE GLOBECOM 2008 Symposium and IEEE ICC 2013 Symposium. He was the founding TPC Co-Chair of CrownCom 2006 and a Steering Committee member of CrownCom 2006–2009. In the area of green communications and networks, he was the lead Guest Editor of the *IEEE Communications Magazine* Feature Topics on Green Communications. He was/is General Co-Chair of IEEE GreenCom 2010 and TPC Co-Chair of IEEE Online GreenComm 2014. He is a Series Editor of *IEEE Communications Magazine* (Green Communications and Computing Networks Series). He is a co-editor/co-author of two books, *Cognitive Communications — Distributed Artificial Intelligence (DAI), Regulatory Policy & Economics, Implementation* (Wiley) and *Green Communications: Theoretical Fundamentals, Algorithms and Applications* (CRC Press).

DANIEL C. KILPER [SM] (dkilper@optics.arizona.edu) is a research professor at the College of Optical Sciences, University of Arizona, and the administrative director of the Center for Integrated Access Networks (www.cianerc.org), an NSF engineering research center. He received his Ph.D. and M.S. in physics from the University of Michigan in 1996 and 1992, respectively, and B.S. degrees in physics and electrical engineering from Virginia Tech in 1994. From 2000 to 2013, he was a member of technical staff at Bell Labs, Alcatel-Lucent. He served as the founding Technical Committee Chair of the GreenTouch Consortium, a global consortium of over 50 organizations, and was the Bell Labs Liaison Executive for the Center for Energy Efficient Telecommunications at the University of Melbourne, Australia. While at Bell Labs, he received the President's Gold Medal Award in 2004 and was a member of the President's Advisory Council on Research. He is an adjunct professor at Columbia University. Currently he is serving as the General Co-Chair of IEEE Online Green Communications Conference 2014 and as TPC Co-Chair for Photonics in Switching 2014. He has conducted research on optical performance monitoring, and on transmission, architectures, and control systems for transparent and energy-efficient optical networks. He holds eight patents, and has authored four book chapters and more than 100 peer-reviewed publications.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE

COMMUNICATIONS EDUCATION AND TRAINING

INDUSTRY CERTIFICATION AND UNIVERSITY ACCREDITATION PROGRAMS

BACKGROUND

Successful design, implementation, maintenance, optimization and improvement of modern communications networks require the efforts of a highly educated, well-trained and dedicated workforce. In recent years, various initiatives have been undertaken which seek to improve our capacity to educate and train both current and future communications workers. These range from new learning technologies and new university accreditation programs to advanced co-operative education programs, training alliances and industry certification programs.

The IEEE Communications Society's Education & Training Board is sponsoring this feature series in order to promote sharing of recent efforts to advance the state of the art in communications education and training. Topics of interest for this edition include recent developments in industry certification and university accreditation programs and their implications.

Our ultimate goal is to recognize innovation in this area and to hasten the adoption of promising new methods and techniques by our community. Original research contributions may also be considered if the authors can present the results in a tutorial fashion that is accessible to non-experts. The submitted materials should not be currently under review by any other journal, magazine or conference.

SUBMISSION GUIDELINES

Prospective authors should follow the IEEE Communications Magazine manuscript format described in the Authors Guidelines (http://www.comsoc.org/commag/paper-submission-guidelines). A typical feature topic consists of 4-6 accepted papers. All articles to be considered for publication must be submitted through the IEEE Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee), according to the following timetable. Select "May 2015/Communications Education" as the category for your submission.

SCHEDULE FOR SUBMISSIONS

Manuscript Submission Due: January 1, 2015
Notification of Acceptance: February 1, 2015
Final Manuscript Due: March 1, 2015
Publication Date: May 2015

GUEST EDITORS

David G. Michelson
Dept. of Electrical and Computer Engineering
The University of British Columbia
Vancouver, Canada
davem@ece.ubc.ca

Tarek El-Bawab
Dept. of Electrical and Computer Engineering
Jackson State University Jackson
Jackson, MS, USA
tarek.el-bawab@jsums.edu

Wen Tong
Wireless CTO
Huawei Technologies
Ottawa, Canada
tongwen@huawei.co

Simultaneous Wireless Information and Power Transfer in Modern Communication Systems

Ioannis Krikidis, Stelios Timotheou, Symeon Nikolaou, Gan Zheng, Derrick Wing Kwan Ng, and Robert Schober

ABSTRACT

Energy harvesting for wireless communication networks is a new paradigm that allows terminals to recharge their batteries from external energy sources in the surrounding environment. A promising energy harvesting technology is wireless power transfer where terminals harvest energy from electromagnetic radiation. Thereby, the energy may be harvested opportunistically from ambient electromagnetic sources or from sources that intentionally transmit electromagnetic energy for energy harvesting purposes. A particularly interesting and challenging scenario arises when sources perform simultaneous wireless information and power transfer (SWIPT), as strong signals not only increase power transfer but also interference. This article provides an overview of SWIPT systems with a particular focus on the hardware realization of rectenna circuits and practical techniques that achieve SWIPT in the domains of time, power, antennas, and space. The article also discusses the benefits of a potential integration of SWIPT technologies in modern communication networks in the context of resource allocation and cooperative cognitive radio networks.

INTRODUCTION

Recently, there has been a lot of interest in integrating energy harvesting technologies into communication networks. Several studies have considered conventional renewable energy resources, such as solar and wind, and have investigated optimal resource allocation techniques for different objective functions and topologies. However, the intermittent and unpredictable nature of these energy sources makes energy harvesting critical for applications where quality of service (QoS) is of paramount importance, and most conventional harvesting technologies are only applicable in certain environments. An energy harvesting technology that overcomes the above limitations is wireless power transfer (WPT), where the nodes charge their batteries from electromagnetic radiation. In WPT, green energy can be harvested from either ambient signals opportunistically or a dedicated source in a fully controlled manner; in the latter case, green energy transfer can take place from more powerful nodes (e.g., base stations) that exploit conventional forms of renewable energy.

Initial efforts on WPT have focused on long-distance and high-power applications. However, both the low efficiency of the transmission process and health concerns with such high-power applications prevented their further development. Therefore, most recent WPT research has focused on near-field energy transmission through inductive coupling (e.g., used for charging cell phones, medical implants, and electrical vehicles). In addition, recent advances in silicon technology have significantly reduced the energy demand of simple wireless devices. WPT is an innovative technology and attracts the interest from both academia and industry; some commercial WPT products already exist (e.g., [1]), and several experimental results for different WPT scenarios are reported in the literature [2]. With sensors and wireless transceivers getting ever smaller and more energy-efficient, we envision that radio waves will not only become a major source of energy for operating these devices, but their information and energy transmission aspects will also be unified. Simultaneous wireless information and power transfer (SWIPT) can result in significant gains in terms of spectral efficiency, time delay, energy consumption, and interference management by superposing information and power transfer. For example, wireless implants can be charged and calibrated concurrently with the same signal, and wireless sensor nodes can be charged with the control signals they receive from the access point. In the era of the Internet of Things, SWIPT technologies can be of fundamental

Ioannis Krikidis and Stelios Timotheou are with the University of Cyprus.

Symeon Nikolaou is with Frederick University.

Gan Zheng is with the University of Essex and the University of Luxembourg.

Derrick Wing Kwan Ng and Robert Schober are with Friedrich-Alexander-University Erlangen-Nurnberg (FAU).

0163-6804/14/$25.00 © 2014 IEEE

importance for energy supply to and information exchange with numerous ultra-low-power sensors, which support heterogeneous sensing applications. Also, future cellular systems with small cells, massive multiple-input multiple-output (MIMO), and millimeter-wave technologies will overcome current path loss effects; in this case, SWIPT could be integrated as an efficient way to jointly support high throughput and energy sustainability.

In this article, we give an overview of the SWIPT technology, and discuss recent advances and future research challenges. More specifically, we explain the rectifying antenna (rectenna) circuit, which converts microwave energy into direct current (DC) electricity and is an essential block for the implementation of WPT/SWIPT technology. Due to practical limitations, SWIPT requires the splitting of the received signal into two orthogonal parts. Recent SWIPT techniques that separate the received signal in the domains of time, power, antenna, and space are presented. On the other hand, SWIPT entails fundamental modifications for the operation of a communication system, and motivates new applications and services. From this perspective, we discuss the impact of SWIPT on the radio resource allocation problem as well as sophisticated cognitive radio (CR) scenarios that enable information and energy cooperation between primary and secondary networks.

WPT Module Components

Exchanging electromagnetic power wirelessly can be classified into three distinct cases:

- Near field power transfer employing inductive, capacitive, or resonant coupling that can transfer power in the range of tenths of Watts, over short distances of up to one meter (sub-wavelength)
- Far field directive power beaming, requiring directive antennas, that can transfer power in the range of several milliWatts at distances of up to several meters in indoor and outdoor environments
- Far field, low-power, ambient RF power scavenging involving receivers that opportunistically scavenge the power transmitted from public random transmitters (cell phone base stations, TV broadcasting stations) for communication with their peer nodes

For this last case the collected power is in the range of several microWatts, and the communication range can be up to several kilometers assuming there is adequate power density. While there are several applications related to near field wireless charging, such as wireless charging of electric cars, cell phones, or other handheld devices, the main focus of this article is on far field WPT, which involves the use of antennas communicating in the far field.

Wireless Power Receiver Module

A wireless power scavenger or receiver consists of the following components: a receiver antenna or antenna array, a matching network, a radio frequency to direct current (RF-DC) converter or rectifier, a power management unit (PMU), and the energy storage unit [3]. Upon the successful charging of the energy storage unit, the storage unit, usually a rechargeable battery or a super capacitor, will provide power to the central processing unit (CPU), the sensors, and the low duty cycle communication transceiver. The schematic of this module is presented in Fig. 1, and a successful implementation of a WPT system that scavenges ambient power 6.3 km away from a Tokyo TV tower is shown in Fig. 2.

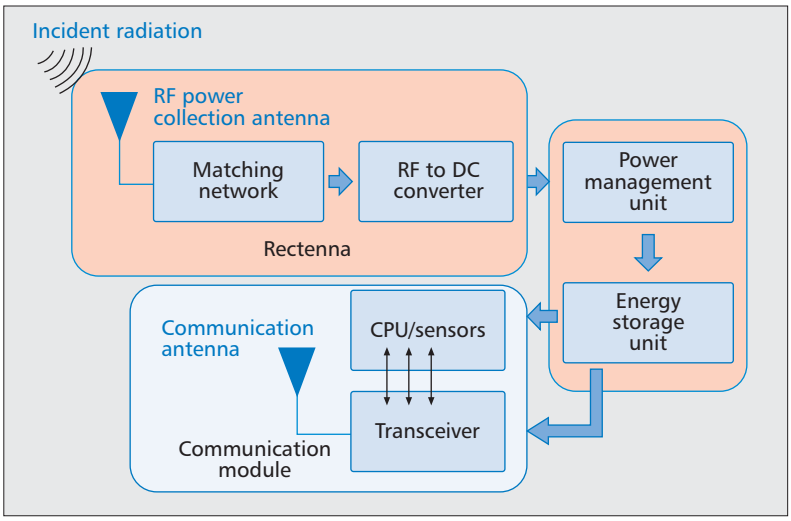


Figure 1. Block diagram of a typical power scavenging module powering a communication transceiver.

Conditions for Efficient WPT

Based on the Friis free space equation, the received RF power at the terminals of the antenna depends on the available power density and the antennas' effective area $A_e = (\lambda^2 G_R)/(4\pi)$ and is given by

$$P_R = \cos^2\phi \frac{P_T G_T}{4\pi R^2} A_e, \quad (1)$$

where P_T and P_R are the transmitted and received power, respectively, G_T and G_R are the transmitter and receiver gains (functions of the spatial variables), respectively, λ denotes the wavelength, and $cos\ \phi$ is the polarization loss factor, which accounts for the misalignment (angle ϕ) of the received electric intensity vector E and the receiver antenna linear polarization vector. From Eq. 1 we can deduce that in order to ensure maximum received power, the receiver antenna needs to have high gain, has to be directed toward the transmitter (maximum directivity direction), and must be aligned with the received E-field ($\phi = 0$). However, these conditions cannot be ensured in practice. For example, in a Rayleigh multipath propagation environment the received signal has random polarization. Consequently, the optimum polarization for a receiver antenna is dual, linear, orthogonal polarization because it ensures the reception of the maximum average power regardless of the received signal's polarization. If the maximum gain direction cannot be guaranteed, omnidirectional antennas are preferred instead. The Friis equation is frequency-depen-

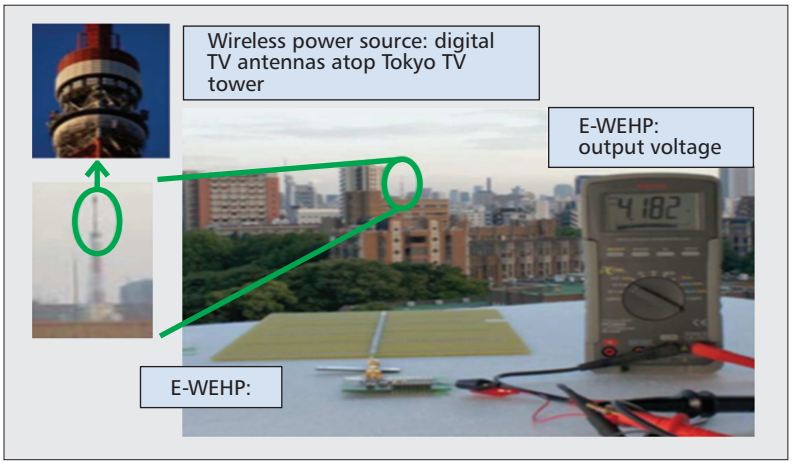


Figure 2. Field measurement in downtown Tokyo, Japan, with prototype device [4] harvesting wireless energy from multicarrier wireless digital TV signals broadcasted from atop the Tokyo TV tower 6:3 km away.

dent and applicable to narrowband signals. The total received power is calculated by integrating the received power P_R over frequency; therefore, a broadband antenna will receive more power than a narrowband one. As a result, wideband antennas or multi-band antennas are preferred.

The RF-to-DC converter or rectifier is probably the most critical component of a WPT module, and its design is the most challenging task [5]. A rectifier consists of at least one nonlinear device. Most rectennas (antenna and rectifier co-design) reported in the literature consist of only one diode. Ideally, the conversion efficiency of a rectifying circuit with a single nonlinear device can reach up to 100 percent. Unfortunately, this can only happen for specific values of P_{RF} and R_{DC}, where P_{RF} denotes the level of the RF power input at the rectifier, and R_{DC} is the delivered load. In more detail, the rectenna structure consists of a single shunt full-wave rectifying circuit with one diode, a $\lambda/4$ distributed line, and a capacitor to reduce the loss in the diode. Depending on the requirements, more complicated and sophisticated rectifier topologies can be used which are based on the well-known Dickson charge pump that can provide both rectification and impedance transformation. Typically, Schottky diodes are used as the nonlinear devices because they have low forward voltage drop and allow very fast switching action, features useful for rectifiers. Low forward voltage drop is needed because the received power is rather small, and fast switching action is needed to follow the relatively high RF frequency of the received signal. Alternatively, it is possible to use complementary metal oxide semiconductor (CMOS) transistors or other transistors as the nonlinear rectifying elements, especially when integrated solutions are preferred. The major problem with RF-to-DC converters is that their efficiency, defined as $n_R = P_{RF}/(V_{DC}^2/R_{DC})$, depends on P_{RF}, R_{DC}, and the DC voltage, V_{DC}, across the load. Generally, the higher the incident RF power, the higher the efficiency. For low power levels, efficiency can even drop to zero because the diodes' forward voltage drop is too high. This is why the reported high efficiencies cannot be seen in actual RF scavenging scenarios. As an example, the ambient power density measured 6.5 km away from the Tokyo TV tower was approximately 1 $\mu W/cm^2$ and the received power was about 50 μW, whereas high efficiency rectifiers require input powers between 0.5–5 mW, 10 to 100 times higher. As a result, the measured efficiency was rather small.

The final stage of the WPT module is the power management unit (PMU), which is responsible for maintaining the optimum load at the terminals of the rectifier despite the changing received RF power levels, and at the same time ensures the charging of the energy storage unit without additional loss.

Techniques for SWIPT

Early information theoretical studies on SWIPT have assumed that the same signal can convey both energy and information without losses, revealing a fundamental trade-off between information and power transfer [10]. However, this simultaneous transfer is not possible in practice, as the energy harvesting operation performed in the RF domain destroys the information content. To practically achieve SWIPT, the received signal has to be split in two distinct parts, one for energy harvesting and one for information decoding. In the following, the techniques that have been proposed to achieve this signal splitting in different domains (time, power, antenna, space) are discussed.

Time Switching

If time switching (TS) is employed, the receiver switches in time between information decoding and energy harvesting [6]. In this case, signal splitting is performed in the time domain, and thus the entire signal received in one time slot is used either for information decoding or power transfer (Fig. 3a). The TS technique allows for a simple hardware implementation at the receiver but requires accurate time synchronization and information/energy scheduling.

Power Splitting

The power splitting (PS) technique achieves SWIPT by splitting the received signal into two streams of different power levels using a PS component; one signal stream is sent to the rectenna circuit for energy harvesting, and the other is converted to baseband for information decoding (Fig. 3b) [6]. The PS technique entails higher receiver complexity compared to TS and requires the optimization of the PS factor α; however, it achieves instantaneous SWIPT, as the signal received in one time slot is used for both information decoding and power transfer. Therefore, it is more suitable for applications with critical information/energy or delay constraints and closer to the information theoretical optimum.

Antenna Switching

Typically, antenna arrays are used to generate DC power for reliable device operation. Inspired by this approach, the antenna switching (AS) technique dynamically switches each antenna element between decoding/rectifying to achieve SWIPT in the antenna domain (Fig. 3c). In the

AS scheme, the receiving antennas are divided into two groups where one group is used for information decoding and the other group for energy harvesting [6]. The AS technique requires the solution of an optimization problem in each communication frame in order to decide the optimal assignment of the antenna elements for information decoding and energy harvesting. For a MIMO decode-and-forward (DF) relay channel, where the relay node uses the harvested energy in order to retransmit the received signal, the optimization problem was formulated as a knapsack problem and solved using dynamic programming in [7].

Because optimal AS suffers from high complexity, low-complexity AS mechanisms have been devised that use the principles of generalized selection combining (GSC) [7]. The key idea of GSC-AS is to use L out of N_T antennas with the strongest channel paths for either energy (GSCE technique) or information (GSCI technique) and the rest for the other operation.

Spatial Switching

The spatial switching (SS) technique can be applied in MIMO configurations and achieves SWIPT in the spatial domain by exploiting the multiple degrees of freedom (DoFs) of the interference channel [8]. Based on the singular value decomposition (SVD) of the MIMO channel, the communication link is transformed into parallel eigenchannels that can convey either information or energy (Fig. 3d). At the output of each eigenchannel there is a switch that drives the channel output to either the conventional decoding circuit or the rectification circuit. Eigenchannel assignment and power allocation in different eigenchannels is a difficult nonlinear combinatorial optimization problem; in [8] an optimal polynomial complexity algorithm has been proposed for the special case of unlimited maximum power per eigenchannel.

Numerical Example — The performance of the discussed SWIPT techniques is illustrated in Fig. 4 for the MIMO relay channel introduced earlier, assuming a normalized block fading Rayleigh. In the considered setup, a single-antenna source communicates with a single-antenna destination through a battery-free MIMO relay node, which uses harvested energy in order to power the relaying transmission. We assume that the source transmits with power P and spectral efficiency r_0 = 2 b/channel use (BPCU); the relay node has global channel knowledge, which enables beamforming for the relaying link. An outage event occurs when the destination is not able to decode the transmitted signal, and the performance metric is the outage probability. The first observation is that GSCI outperforms GSCE scheme for $L = 1$ and $L = 2$, respectively. This result shows that diversity gain becomes more important than energy harvesting due to the high RF-to-DC efficiency η. In addition, the GSCI scheme with $L = 1$ is the optimal GSC-based strategy and achieves a diversity gain equal to two. It can also be seen that the PS scheme outperforms the AS scheme with a gain of 2.5 dB for high P, while the TS scheme provides poor performance due to the required time division.

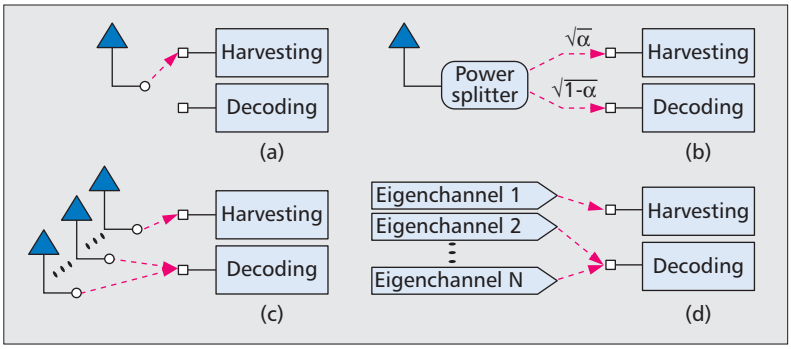


Figure 3. SWIPT transmission techniques in different domains: a) time; b) power; c) antenna; d) space.

Resource Allocation for Systems with SWIPT

This section discusses the benefits of employing SWIPT on resource allocation applications. Utility-based resource allocation algorithm design has been heavily studied in the literature [9] for optimizing the utilization of limited resources in the physical layer such as energy, bandwidth, time, and space in multiuser systems. In addition to the conventional QoS requirements such as throughput, reliability, energy efficiency, fairness, and delay, the efficient transfer of energy plays an important role as a new QoS requirement for SWIPT [10, 11]. Resource allocation algorithm design for SWIPT systems includes the following aspects:

Joint power control and user scheduling: The RF signal acts as a dual-purpose carrier for conveying information and energy to the receivers simultaneously. However, the wide dynamic range of the power sensitivity for energy harvesting (–10 dBm) and information decoding (–60 dBm) is an obstacle to realizing SWIPT. As a result, joint power control and user scheduling is a key aspect for facilitating SWIPT in practice. For instance, idle users experiencing high channel gains can be scheduled for power transfer to extend the lifetime of the communication network. Besides, opportunistic power control can be used to exploit the channel fading for improved energy and information transfer efficiency [10–12]. Figure 5 depicts an example of power control in SWIPT systems. We show the average system capacity vs. the average total harvested energy in a downlink system. In particular, a transmitter equipped with N_T antennas is serving one single-antenna information receiver and K single-antenna energy harvesting receivers. As can be observed, with optimal power control, the trade-off region of the system capacity and the harvested energy increases significantly with N_T. Besides, the average harvested energy improves with the number of energy harvesting receivers.

Energy and information scheduling: For passive receivers such as small sensor nodes, uplink data transmission is only possible after the receivers have harvested a sufficient amount of energy from the RF in the downlink. The physical constraint on the energy usage motivates a "harvest-then-transmit" design. Allocating more

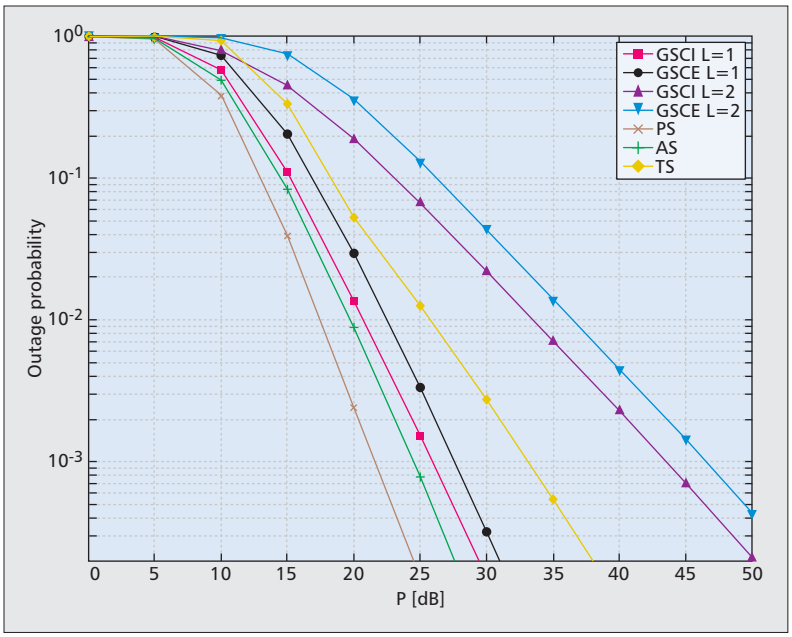


Figure 4. Outage probability versus P for GSCI, GSCE, PS, AS and TS; the simulation setup is $r_0 = 2$ BPCU, $N_T = 3$ antennas, $L = \{1, 2\}$ and RF-to-DC efficiency $\eta = 1$.

time for energy harvesting in the downlink leads to a higher amount of harvested energy, which can then be used in the uplink. However, this also implies that there is less time for uplink transmission which may result in a lower transmission data rate. Thus, by varying the amounts of time allocated for energy harvesting and information transmission, the system throughput can be optimized.

Interference management: In traditional communication networks, co-channel interference is recognized as one of the major factors that limits the system performance and is suppressed or avoided via resource allocation. However, in SWIPT systems, the receivers may embrace strong interference since it can act as a vital source of energy. In fact, injecting artificial interference into the communication network may be beneficial for the overall system performance, especially when the receivers do not have enough energy for supporting their normal operations, since in this case, information decoding becomes less important compared to energy harvesting. Besides, by exploiting interference alignment and/or interference coordination, a "wireless charging zone" can be created by concentrating and gathering multicell interference in certain locations.

Joint Information and Energy Cooperation in CR Networks

SWIPT also opens up new opportunities for cooperative communications. We present one example where SWIPT improves the traditional system design of cooperative CR networks (CCRNs). CCRNs are a new paradigm for improving spectrum sharing by having the primary and secondary systems actively seek opportunities to cooperate with each other [13, 14]. The secondary transmitter (ST) helps relay the traffic of the primary transmitter (PT) to the primary user (PU), and in return can utilize the primary spectrum to serve its own secondary user (SU). However, to enable this cooperation, the ST should both possess a good channel link to the primary system and have sufficient transmit power. While the former can be achieved by proper placement, the latter requirement cannot be easily met especially when the ST is a low-power relay node rather than a powerful base station (BS), which renders this cooperation unmeaningful.

SWIPT may provide a promising solution to address this challenge by encouraging the cooperation between the primary and secondary systems at both the information and energy levels [15]. That is, the PT will transmit both information and energy to the ST; in exchange, the low-power ST relays the primary information. Compared to the traditional CCRN, this approach creates more incentives for both systems to cooperate and therefore improves the system overall spectrum efficiency without relying on external energy sources.

We illustrate the performance gain by studying a joint information and energy cooperation scheme using the amplify-and-forward protocol and the power splitting technique. Two channel phases are required to complete the communication. In phase I, the PT broadcasts its data and both the ST and the PU listen. The ST then splits the received RF signal into two parts: one for information processing then forwarding to the PU, and the other for harvesting energy, with relative power ratio of α and $1 - \alpha$,respectively. In phase II, the ST superimposes the processed primary data with its own precoded data, then transmits it to both the PU and the SU. The ST jointly optimizes power allocation factor α and the precoding vectors to the PU and SU to achieve the maximum rates.

In Fig. 6, we show the achievable rate region of the proposed information and energy cooperation schemes, and compare it with the conventional information cooperation only scheme [14]. We consider a scenario where the distances from the ST to all the other terminals are 1 m, while the distance from the PT to the PU is 2 m, and thus assistance from the ST is usually preferred by the PT. We assume that the ST has four transmit antennas and all other terminals have a single antenna. The primary energy is set to 20 dB, while the available secondary energy is 10 dB. The path loss exponent is 3.5, and the K factor for the Rician channel model is set to 5 dB. The RF-to-DC efficiency is equal to $\eta = 0.1$, 0.5, and 1. It is seen that the achievable rate regions are greatly enlarged thanks to the extra energy cooperation even with RF-to-DC efficiencies as low as $\eta = 0.1$. When the required PU rate is 2 b/s/Hz, the SU can double or triple its rate compared to the case without energy cooperation as η varies from 0.1 to 1. When the SU rate is 1.5 b/s/Hz, the PU enjoys a 75 percent higher data rate when $\eta = 1$. The proposed additional energy cooperation clearly introduces a substantial performance gain over the existing information cooperation only CR scheme, and could be a promising solution for future CCRNs.

Conclusion and Future Work

This survey provides an overview of SWIPT technology. Different SWIPT techniques that split the received signal in orthogonal components have been discussed. We have shown that SWIPT introduces fundamental changes in the resource allocation problem and influences basic operations such as scheduling, power control, and interference management. Finally, a sophisticated CR network that enables information/energy cooperation between primary and secondary systems has been discussed as an example of new SWIPT applications. SWIPT imposes many interesting and challenging new research problems and will be a key technology for the next-generation communication systems. In the following, we discuss some of the research challenges and potential solutions.

Path loss: The efficiency of SWIPT is expected to be unsatisfactory for long-distance transmission unless advanced resource allocation and antenna technology can be combined. Two possible approaches to overcome this problem include the use of massive MIMO and coordinated multipoint systems. The former increase the DoF offered to harvest energy and create highly directive energy/information beams steered toward the receivers. The latter provides spatial diversity for combating path loss by reducing the distance between transmitters and receivers. Besides, the distributed transmitters may be equipped with traditional energy harvesters (e.g., solar panels) and exchange their harvested energy over a power grid to overcome potential energy harvesting imbalances in the network.

Communication and energy security: Transmitters can increase the energy of the information carrying signal to facilitate energy harvesting at the receivers. However, this may also increase their susceptibility to eavesdropping due to the broadcast nature of wireless channels. On the other hand, receivers requiring power transfer may take advantage of the transmitter by falsifying their reported channel state information. Therefore, new QoS concerns regarding communication and energy security naturally arise in SWIPT systems.

Hardware development: Despite the wealth of theoretical techniques for SWIPT, so far, hardware implementations have mostly been limited to WPT systems that opportunistically harvest ambient energy. Thus, the development of SWIPT circuits is fundamental to investigate the trade-off between SWIPT techniques occurring due to inefficiencies of different circuit modules. For example, the TS technique is theoretically less efficient than PS, but the latter suffers from power splitting losses that are not accounted for in theoretical studies.

Applications: SWIPT technology has promising applications in several areas that can benefit from ultra-low-power sensing devices. Potential applications include structure monitoring by embedding sensors in buildings, bridges, roads, and so on; healthcare monitoring using implantable bio-medical sensors; and building automation through smart sensors that monitor and control different building processes. However, for successful realization of such SWIPT applications, several challenges must be overcome at various layers from hardware implementation over protocol development to architectural design.

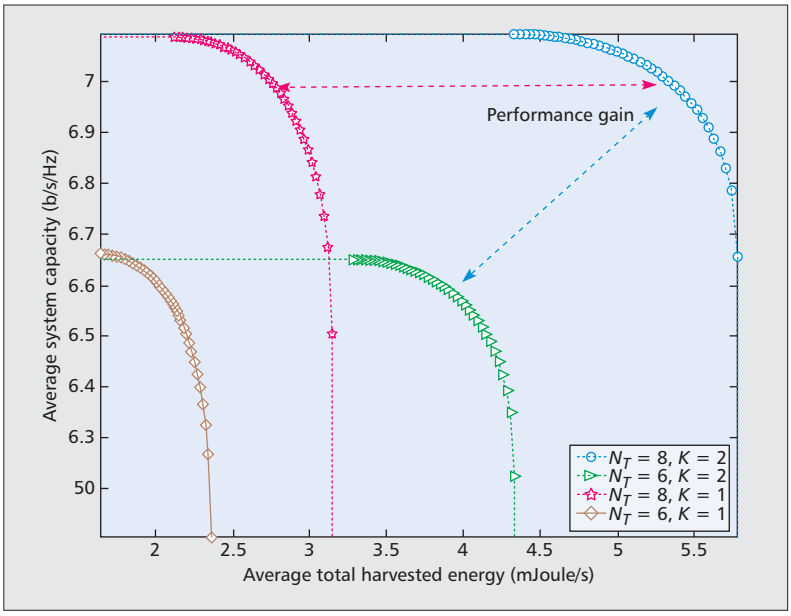


Figure 5. The trade-off region of the average system capacity (b/s/Hz) and the average total harvested energy (mJ/s) for different numbers of receivers. The carrier frequency is 915 MHz, and the information receiver and energy harvesting receivers are located at 30 m and 10 m from the transmitter, respectively. The total transmit power, noise power, transceiver antenna gain, and RF-to-DC conversion loss are set to 10 W, –23 dBm, 10 dBi, and 3 dB, respectively.

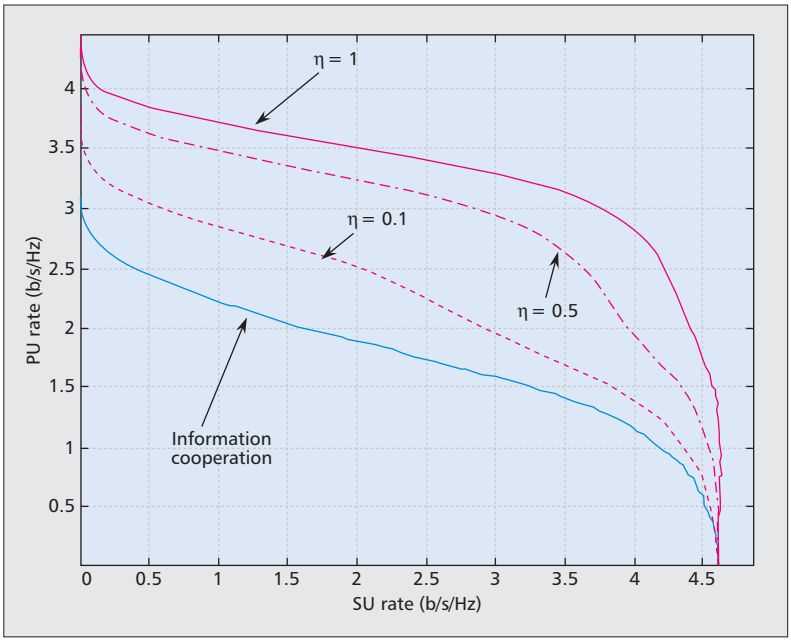


Figure 6. PU-SU rate region with different values of RF-DC efficiency.

Acknowledgment

This work was partially supported by the Research Promotion Foundation, Cyprus, under the project KOYLTOYRA/BP-NE/0613/04, "Full-Duplex Radio: Modeling, Analysis and Design (FD-RD)."

SWIPT technology has promising applications in several areas that can benefit from ultra-low power sensing devices. However, for the successful realization of such SWIPT applications, several challenges have to be overcome at various layers.

References

[1] www.powercastco.com.
[2] N. Shinohara, "Development of Rectenna with Wireless Communication System," *Proc. Euro. Conf. Ant. Prop.*, Rome, Italy, Apr. 2011, pp. 3970–73.
[3] Z. Popovic, "Cut the Cord: Low-Power Far-Field Wireless Powering," *IEEE Microwave Mag.*, vol. 14, 2013, pp. 55–62.
[4] R. J. Vyas *et al.*, "E-WEHP: A Batteryless Embedded Sensor Platform Wirelessly Powered from Ambient Digital-TV Signal," *IEEE Trans. Microwave Theory and Tech.*, vol. 61, June 2013, pp. 2491–2505.
[5] P. Nintanavongsa *et al.*, "Design Optimization and Implementation for RF Energy Harvesting Circuits," *IEEE J. Emerging Sel. Topics Circuit Sys.*, vol. 2, Mar. 2012, pp. 24–33.
[6] R. Zhang and C. K. Ho, "MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 12, May 2013, pp. 1989–2001.
[7] I. Krikidis *et al.*, "A Low Complexity Antenna Switching for Joint Wireless Information and Energy Transfer in MIMO Relay Channels," *IEEE Trans. Commun.*, vol. 62, no. 5, May 2014, pp. 1577–87.
[8] S. Timotheou and I. Krikidis, "Joint Information and Energy Transfer in the Spatial Domain with Channel Estimation Error," *Proc. IEEE Online Conf. Green Commun.*, Oct. 2013, pp. 115–20.
[9] G. Song and Y. Li, "Utility-based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Networks," *IEEE Commun. Mag.*, vol. 43, Dec. 2005, pp. 127–34.
[10] P. Grover and A. Sahai, "Shannon Meets Tesla: Wireless Information and Power Transfer," *Proc. IEEE Int'l. Symp. Info. Theory*, June 2010, Austin, TX, pp. 2363–67.
[11] D. W. K. Ng, E. S. Lo, and R. Schober, "Wireless Information and Power Transfer: Energy Efficiency Optimization in OFDMA Systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, Dec. 2013, pp. 6352–70.
[12] D. W. K. Ng, E. S. Lo, and R. Schober, "Robust Beamforming for Secure Communication in Systems with Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 13, Aug. 2014, pp. 4599–4615.
[13] O. Simeone *et al.*, "Spectrum Leasing to Cooperating Secondary Ad Hoc Networks," *IEEE JSAC*, vol. 26, no. 1, Jan. 2008, pp. 203–13.
[14] G. Zheng *et al.*, "Cooperative Cognitive Networks: Optimal, Distributed and Low-Complexity Algorithms," *IEEE Trans. Signal Proc.*, vol. 61, no. 11, June 2013, pp. 277–90.
[15] G. Zheng *et al.*, "Information and Energy Cooperation in Cognitive Radio Networks," *IEEE Trans. Sig. Proc.*, vol. 62, no. 9, May 2014, pp. 2290–2303.

Biographies

IOANNIS KRIKIDIS [S'03, M'07, SM'12] (krikidis@ucy.ac.cy) received his diploma in computer engineering from the Computer Engineering and Informatics Department of the University of Patras, Greece, in 2000, and the M.Sc and Ph.D degrees from Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2001 and 2005, respectively, all in electrical engineering. From 2006 to 2007 he worked as a post-doctoral researcher at ENST, and from 2007 to 2010 he was a research fellow in the School of Engineering and Electronics at the University of Edinburgh, United Kingdom. He is currently an assistant professor at the Department of Electrical and Computer Engineering, University of Cyprus, Nicosia. His current research interests include information theory, wireless communications, cooperative communications, cognitive radio, and secrecy communications.

STELIOS TIMOTHEOU [S'04, M'10] (timotheou.stelios@ucy.ac.cy) received a B.Sc. from the Electrical and Computer Engineering School of the National Technical University of Athens, and an M.Sc. and a Ph.D. from the Electrical and Electronic Engineering Department of Imperial College London. He is currently a research associate at the KIOS Research Center for Intelligent Systems and Networks of the University of Cyprus. His research focuses on the modeling and system-wide solution of problems in complex and uncertain environments that require real-time and close to optimal decisions by developing optimization, machine learning, and computational intelligence techniques.

SYMEON NIKOLAOU [S'04, M'07] (eng.ns@frederick.ac.cy) received his Diploma degree in electrical and computer engineering (Magna Cum Laude) from the National Technical University of Athens in 2003, and his M.Sc. and Ph.D. degrees in electrical and computer engineering from Georgia Institute of Technology, Atlanta in 2005 and 2007, respectively. Since 2007 he has been a faculty member at Frederick University and a senior researcher at Frederick Research Center (FRC). He is a member of the Microwaves and Photonics research group, and his research interests span UWB antennas, passive and active RFIDs, wireless sensors applications, sensor integration with RFIDs, wearable and reconfigurable antennas, microwave imaging using UWB sensor arrays, and wireless power transfer.

GAN ZHENG [S'05, M'09, SM'12] (ganzheng@essex.ac.uk) received his B. Eng. and M. Eng. from Tianjin University, China, in 2002 and 2004, respectively, both in electronic and information engineering, and his Ph.D. degree in electrical and electronic engineering from the University of Hong Kong in 2008. He worked as a research associate at University College London and the University of Luxembourg during December 2007–September 2010 and September 2010–August 2013, respectively. He is currently a lecturer in the School of Computer Science and Electronic Engineering, University of Essex. He is also affiliated with the University of Luxembourg. His research interests are in the general area of signal processing for wireless communications with current emphasis on cooperative communications, cognitive radio, multi-cell cooperation, physical-layer security, full-duplex radio, and energy harvesting.

DERRICK WING KWAN NG [S'06, M'12] (kwan@lnt.de) received his Bachelor's degree with first class honors and M.Phil. degree in electronic engineering from the Hong Kong University of Science and Technology in 2006 and 2008, respectively. He received his Ph.D. degree from the University of British Columbia in 2012. In the summer of 2011 and spring of 2012, he was a visiting scholar at the Centre Tecnológic de Telecomunicacions de Catalunya — Hong Kong (CTTC-HK). He is now working as a postdoctoral fellow in the Institute for Digital Communications, Friedrich Alexander University Erlangen-Nürnberg, Germany. His research interests include cross-layer optimization for wireless communication systems, resource allocation in OFDMA wireless systems, and communication theory.

ROBERT SCHOBER [S'98, M'01, SM'08, F'10] (schober@lnt.de) received his Diplom (Univ.) and Ph.D. degrees in electrical engineering from the University of Erlangen-Nürnberg in 1997 and 2000, respectively. Since May 2002 he has been with the University of British Columbia (UBC), Vancouver, Canada, where he is now a full professor. Since January 2012 he is an Alexander von Humboldt Professor and the Chair for Digital Communication at Friedrich Alexander University. His research interests fall into the broad areas of communication theory, wireless communications, and statistical signal processing. He received several awards for his work including the 2002 Heinz Maier-Leibnitz Award of the German Science Foundation, the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, a 2011 Alexander von Humboldt Professorship, and a 2012 NSERC E.W.R. Steacie Fellowship. He is currently the Editor-in-Chief of *IEEE Transactions on Communications*.

Green Transmission Technologies for Balancing the Energy Efficiency and Spectrum Efficiency Trade-off

Yiqun Wu, Yan Chen, Jie Tang, Daniel K. C. So, Zhikun Xu, Chih-Lin I, Paul Ferrand, Jean-Marie Gorce, Chih-Hsuan Tang, Pei-Rong Li, Kai-Ten Feng, Li-Chun Wang, Kai Börner, and Lars Thiele

ABSTRACT

As 4G wireless networks are vastly and rapidly deployed worldwide, 5G with its advanced vision of all connected world and zero distance communications is already at the corner. Along with the super quality of user experience brought by these new networks, the shockingly increasing energy consumption of wireless networks has become a worrying economic issue for operators and a big challenge for sustainable development. Green Transmission Technologies (GTT) is a project focusing on the energy-efficient design of physical-layer transmission technologies and MAC-layer radio resource management in wireless networks. In particular, fundamental trade-offs between spectrum efficiency and energy efficiency have been identified and explored for energy-efficiency-oriented design and optimization. In this article, four selected GTT solutions are introduced, focusing on how they utilize the degrees of freedom in different resource domains, as well as how they balance the trade-off between energy and spectrum efficiency. On top of the elaboration of separated solutions, the GTT toolbox is introduced as a systematic tool and unified simulation platform to integrate the proposed GTT solutions together.

INTRODUCTION

The energy consumption of information and communications technology (ICT) has recently become an economic issue for operators as well as a big challenge for sustainable development. The energy consumption of the ICT industry contributes about 3 percent to the global annual electricity bill, and the amount is rising at the speed of 15–20 percent each year [1]. With vast and rapid deployment of fourth generation (4G) networks, as well as the 5G vision of a totally connected world, with zero waiting and zero distance communications by 2020 [2], the situation is getting worse. Motivated by these facts, the GreenTouch Consortium has been founded and aims to improve end-to-end energy efficiency by 1000 fold.[1] Green Transmission Technology (GTT), one of the biggest umbrella projects in GreenTouch, focuses on the energy-efficient design of physical layer transmission technologies and medium access control (MAC) layer radio resource management in wireless networks.

Extensive research has been carried out in the literature on energy-efficient wireless networks, and diverse technologies have been proposed in all aspects, trying to close the gap between practice and expectations [3–8]. To develop a systematic way to evaluate diverse technologies and find the remaining gaps for further optimization, a unified framework is needed. This is the fundamental framework of the energy efficiency (EE) and spectrum efficiency (SE) trade-off [9], which has long been pointed out by Shannon's ground-breaking theory but has yet to be fully utilized. A widely accepted definition of EE is transmitted bits per unit energy, and SE is usually defined as transmission rate per unit bandwidth. In the case of an additive white Gaussian noise (AWGN) channel, the channel capacity is given by Shannon's formula,

$$C = W \log_2\left(1+\frac{P_t G}{WN_0}\right),$$

where W is the bandwidth, P_t is the transmit power, G is the channel gain, and N_0 is the power spectral density of noise. In this case, it is shown that EE is a monotonic decreasing function of SE [4], which implies that optimizing with respect to only one metric, EE or SE, tends to degrade the other, SE or EE.

The reason that the EE-SE trade-off framework is not widely used as expected for the design of green transmission technologies may lie in the fact that practical systems behave differently than AWGN channels. Challenges may

Yiqun Wu and Yan Chen are with Huawei Technologies.

Jie Tang and Daniel K. C. So are with the University of Manchester.

Zhikun Xu and Chih-Lin I are with China Mobile.

Paul Ferrand and Jean-Marie Gorce are with, INSA-Lyon.

Chih-Hsuan Tang is with Chunghwa Telecom.

Pei-Rong Li, Kai-Ten Feng, and Li-Chun Wang are with National Chiao Tung University,

Kai Börner and Lars Thiele are with Fraunhofer HHI.

This article belongs to the joint work of the Green Transmission Technology project in GreenTouch.

[1] For more information about GreenTouch please refer to www.greentouch.org.

0163-6804/14/$25.00 © 2014 IEEE

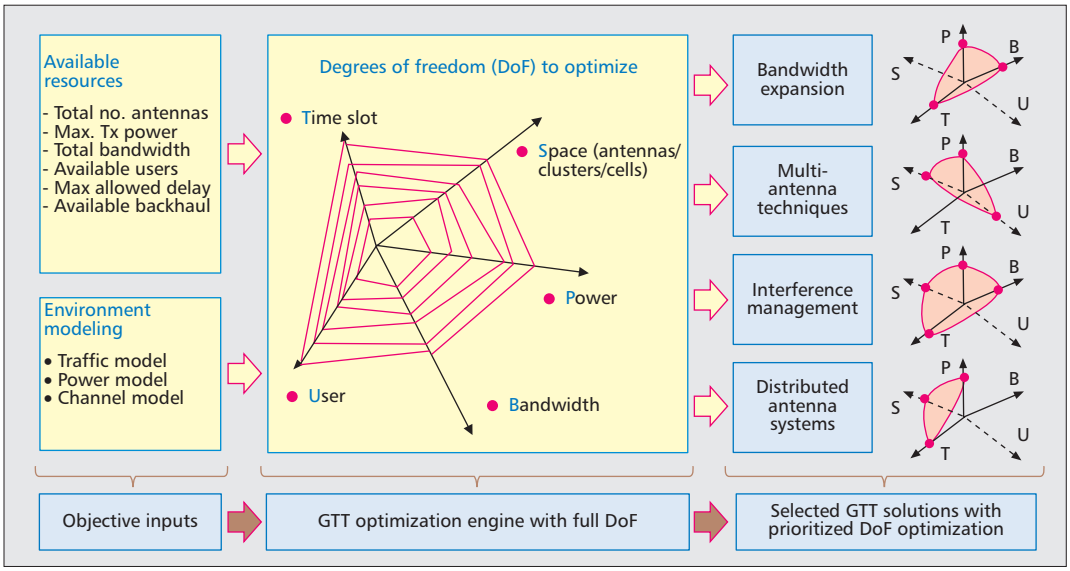


Figure 1. Overview of the design principle and optimization framework of the GTT project.

Our findings show that by the deployment of DAS, or known as coordinated RAN, site power can be greatly reduced, and with the smart use of the DoF in the three domains, the network energy efficiency can be improved over 300 percent.

arise from multi-path fading channels, multi-user sharing the same radio resources (time, frequency, power, antenna, etc.), intercell interference, and the real power consumption model beyond only transmit power. The joint impact of all these factors together usually results in multi-objective and complex optimization problems to derive the EE-SE trade-off. One simple example about the impact of approximate power model is given below. Inside the wireless transceivers, there are various components contributing to the total power consumption, including power amplifier, signal processing circuits, power supply, cooling, and so on. One popular approach is to model the total power consumption as a linear function of transmit power [10], $P_{tot} = \alpha P_t + P_0$, where α is a constant, and P_0 is the static power. With the total power consumption taken into consideration, the curve of EE-SE relation turns to a bell shape [9] (i.e., partially increases and then decreases). Hence, there is an optimal operational point in terms of EE.

Figure 1 gives an overview of the design principle and optimization framework in the GTT project. For each specific network scenario, the available resources and environmental models are identified as objective inputs. The network performance is optimized with all the manageable degrees of freedom (DoFs) as the optimization variables, which include power, time, space, bandwidth, and user. Inspired by the EE-SE trade-off framework, various GTT solutions are proposed, and each GTT solution has its own prioritized DoF optimization. It should be noted that this article is not a dedicated survey of EE-SE trade-off; only the following four selected GTT solutions will be introduced.

Bandwidth expansion, in which the optimization of EE-SE trade-off will focus on the prioritized DoF in radio resources such as power, bandwidth, and time. A new insight here is that to strike a good balance between EE and SE, it is not wise to utilize the radio resources to their extreme ends (as in the case of optimizing SE only). On the contrary, with optimal design, some extra sacrifice in bandwidth (power) could bring substantial savings in power (bandwidth), which is a good bargain.

Multiple antenna techniques, in which the prioritized DoF for the optimization of EE-SE trade-off will be power, space (subspaces created by multi-antenna schemes), and available users (scheduled and grouped for simultaneous service by multiple antennas). Our research results show that given a large number of antennas, simultaneously serving multiple users with multi-streams helps to increase both EE and SE, which may not hold for the case of a small number of antennas.

Intercell interference management, in which the prioritized DoF for EE-SE optimization will be time, power, bandwidth, and space (cells or subspaces created by intercell coordinated transmission techniques). As is pointed out and theoretically verified, it is crucial to cancel at least the strongest interference, which could bring over 10 dB gain in useful signal strength and result in order-wise EE improvement for a given SE requirement.

Distributed antenna system, in which the prioritized DoF for EE-SE optimization will be power, time, and space (collaborative clusters created by joint transmission of several antennas). Our findings show that by the deployment of a distributed antenna system (DAS), known as a coordinated radio access network (RAN), site power can be greatly reduced, and with the smart use of the DoF in the three domains, network EE can be improved over 300 percent.

How these solutions are motivated by the EE-SE trade-off framework and how they may utilize different sets of prioritized DoF to balance the EE-SE trade-off are further elaborated later.

For each GTT solution, the EE-SE trade-off is optimized with its own prioritized DoF. However, as shown in Fig. 1, the prioritized DoF of different GTT solutions may overlap, so these solutions are usually interrelated, with some of them even competing with each other. As a result, even with clear understanding of how each solution makes the best trade-off between

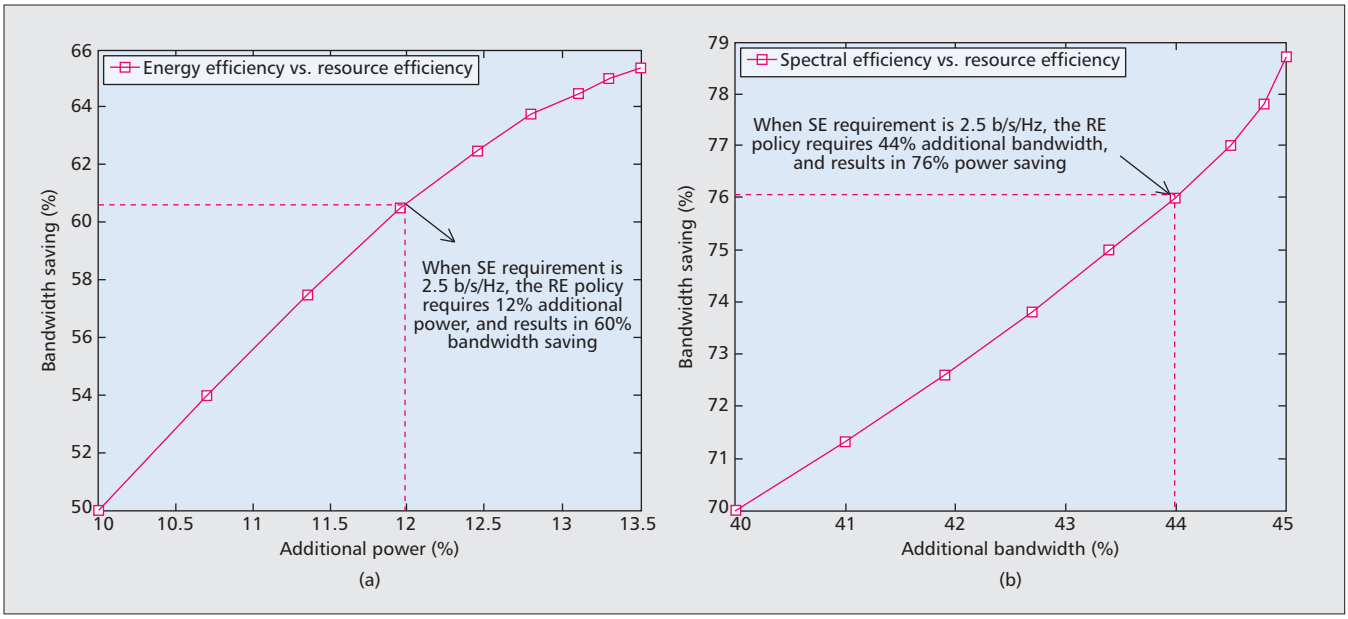


Figure 2. Performance of the RE policy compared to the best EE policy and the best SE policy, and each point corresponds to different minimum SE requirements (1 b/s/Hz : 0.5 b/s/Hz : 5 b/s/Hz), $\gamma = 1$: a) EE vs. RE for a trade-off between extra power and bandwidth saving; b) SE vs. RE for trade-off between extra bandwidth and power saving.

EE and SE, it is still uncertain and worth investigating how much EE can be improved with the joint efforts of all the solutions, and how to make the best trade-off between EE and SE in the presence of all these solutions. Therefore, beyond investigation of the separate solutions, the GTT toolbox is introduced as a systematic and unified simulation platform to integrate all these GTT solutions together, the key methodology of which is then elaborated on later. Finally, we summarize the key findings and conclude the entire article.

Green Transmission Technologies Inspired by EE-SE Framework

Bandwidth Expansion

Based on Shannon's formula, expanding the transmit bandwidth reduces the transmit power under the same rate requirement. In other words, under the framework of EE-SE trade-off, bandwidth expansion increases EE and decreases SE. As available spectrum is limited and expensive in practical systems, traditional design of mobile wireless networks mainly focuses on how to optimally utilize the available spectrum. However, the traffic load in wireless networks has significant spatial and temporal fluctuations due to user mobility and the bursty nature of data applications. During the peak hours, there are a lot of users waiting to be scheduled, and the bandwidth allocated to each user may not be expandable. On the contrary, during the off-peak hours, bandwidth expansion can be applied to improve EE.

Rather than solving the EE-SE trade-off problem for a single-point solution (i.e., a pair of EE and SE outputs), a new approach is proposed by GTT for EE-SE trade-off, the resource efficiency (RE) metric for wireless networks, which is defined as follows:

$$\eta_{RE} = \eta_{EE} + \bar{\gamma}\eta_{SE} \text{ [11]},$$

where

$$\bar{\gamma} = \gamma \frac{W_{tot}}{P_{tot}}.$$

This is a multi-objective optimization problem. There is no a priori correspondence between a weight vector and a solution vector, and it is up to the operator to choose appropriate weights. In particular, γ is used to balance between EE and SE; hence, appropriate weights need to be chosen to utilize the available power and bandwidth. Simulation results in Fig. 2 show that a significant amount of bandwidth can be saved with a slight increase in energy consumption, and a similar conclusion can also be drawn on energy saving by bandwidth expansion. More crucially, it shows that by operating at a slightly reduced EE or SE, the amount of bandwidth or power saved can be substantial. These saved resources can be utilized in other ways; for instance, the vacant bandwidth can be used for improved interference avoidance in a heterogeneous network scenario, or for a system with cognitive radio subsystems. Hence, the available network resources can be utilized efficiently by designing based on the RE metric.

In the multi-cell scenario, if all the cells employ bandwidth expansion, the transmit power is reduced, and the intercell interference is further reduced. Thus, the performance gain of bandwidth expansion can be more in the multi-cell scenario compared to the single-cell scenario. Our recent work in [12] shows that bandwidth expansion enables savings of power consumption of up to 45 percent if all cells in

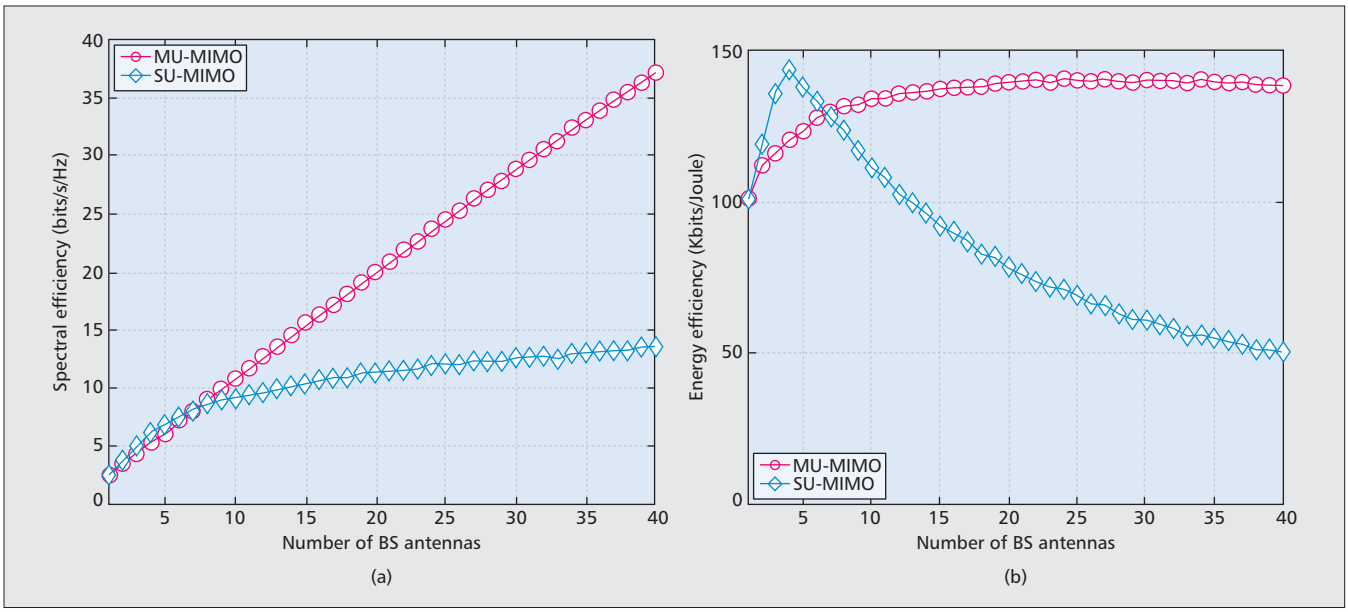


Figure 3. Performance comparison of SU-MIMO and MU-MIMO in a single-cell scenario: a) SE vs. the number of antennas; b) EE vs. the number of antennas.

the network apply the same bandwidth expansion strategy.

Therefore, the potential solution of bandwidth expansion is to jointly optimize the operations of user scheduling, power, and bandwidth allocation in order to maximize the RE for complete wireless networks.The future work for bandwidth expansion may include:

Joint time-frequency expansion: Previous work on bandwidth expansion mainly explores DoFs in the frequency domain. The idea can also be extended to the time domain. The transmit power can be reduced by expanding the transmit time. On the other hand, when the transmitters are idle, they can be put into sleep mode to save energy. Consequently, joint time and frequency domain optimization is a promising way to save energy.

Overhead issues: Bandwidth expansion may not always be beneficial for EE in practice because when transmit bandwidth increases, the overhead such as pilots for channel estimation will increase, and the power consumption for signal processing may also increase. Thus, more efforts are needed to investigate the benefits of bandwidth expansion in practice, including developing new power models and new strategies for bandwidth expansion.

Multiple-Antenna Techniques

Multiple-antenna techniques play an important role in wireless networks today. If multiple antennas are applied at both the transmitter and receiver, it can be regarded as a multiple-input multiple-output (MIMO) system. Additional spatial DoFs by applying multiple antennas enhance the reliability and significantly increase the transmission rate, without additional bandwidth or power. Thus, the EE-SE trade-off can be improved by multiple-antenna technologies.

Multiple-antenna technologies can reduce the transmit power by several means. One approach is beamforming, by which signals can be combined constructively at the receiver. The power gain of beamforming is $n_t n_r$, where n_t is the number of transmit antennas and n_r is the number of receive antennas. Transmit power can also be reduced by spatial diversity schemes, which provide redundancy across independent fading branches. Given the same outage probability, much less power is required by spatial diversity schemes. Another way to reduce the transmit power is spatial multiplexing, by which multiple data streams can be transmitted in parallel. Given the same rate requirement, the required transmit power can be reduced.

In cellular networks, base stations (BSs) have more antennas than users, and the MIMO channel capacity is limited by the minimum of n_t and n_r. In this case, multi-user MIMO (MU-MIMO) can be applied to improve the system capacity. We compare the EE and SE performance of a MU-MIMO scheme and a single-user MIMO (SU-MIMO) scheme in a single-cell scenario by simulation in Fig. 3. In the simulation, zero-forcing precoding is applied for both schemes, and the number of scheduled users is the maximum value for MU-MIMO. The results show that if the number of antennas is small, SU-MIMO outperforms MU-MIMO in both SE and EE. As the number of antennas increases, the advantage of MU-MIMO over SU-MIMO is enlarged. Therefore, it is beneficial to switch between MU-MIMO and SU-MIMO for better EE-SE trade-off in practice.

Multiple-antenna techniques also have detrimental effects on energy consumption. First, more circuit energy is consumed for MIMO transmission as additional RF chains are required, and the complexity of signal processing is also higher. Second, more time or frequency resources are spent on the signaling overhead for MIMO transmission. For example, channel state information (CSI) is required for detection at the receiver and precoding at the transmitter. To estimate the CSI at the receiver and feed it

Reducing interference can be achieved by a proper static or dynamic resource allocation over cells providing a significant EE gain. However, most of these techniques rely implicitly on reducing the bandwidth available for each BS thus introducing a SE loss, and the global trade-off is then shifted.

back to the transmitter, training symbols, pilots, and control signals are transmitted. Since the number of channel coefficients increases with n_t n_r, much more signaling overhead is required for MIMO systems, especially when there are a large number of antennas and users.

MIMO transmission schemes reduce the transmit power but increase the circuit power, so the total power consumption may not always be reduced. If the increase of transmit rate cannot compensate the increase of total power, EE will decrease. To improve EE, the above benefits and detriments need to be balanced. One approach is to adaptively turn on/off the antennas and related RF chains so that the EE and SE can be well balanced. In general, the EE-SE trade-off problem with multiple-antenna techniques is complicated, involving joint optimization of precoding, scheduling, power allocation, and antenna selection. The future work for multiple antenna technologies may include:

Transmission scheme adaptation: There are many different MIMO transmission schemes, such as spatial diversity, beamforming, multiplexing, and MU-MIMO. To achieve the best performance, the transmission schemes need to be adaptively selected based on the channel state information (CSI) and user requirement.

Sleep mode management: If spectral efficiency can be improved by MIMO transmission, the transmission time is less, and the transmitter can be put into sleep mode. On the other hand, the number of active antennas for each transmitter can also be adaptive. Therefore, joint node-level and antenna-level sleep mode management will be interesting work.

CSI acquisition: CSI plays an important role in multiple-antenna techniques. Without accurate CSI, performance will be largely degraded. The accuracy of CSI can be improved if more signaling is done. How to balance the trade-off between CSI accuracy and signaling overhead is valuable and critical work, and needs to be further investigated.

Intercell Interference Management

The fundamental challenge for a multi-cell scenario is the mitigation of intercell interference, especially when the frequency is in full reuse. The EE and SE will be significantly degraded by intercell interference, especially for cell edge users. Reducing interference can be achieved by proper static or dynamic resource allocation over cells providing a significant EE gain. However, most of these techniques rely implicitly on reducing the bandwidth available for each BS, thus introducing an SE loss, and the global trade-off is then shifted.

In GTT we analyzed the global impact of intercell interference, which, in degrading both EE and SE, results in poorer system performance. Two extreme cases are evaluated on a reference scenario corresponding to a current dense urban standard deployment under full load. The first case corresponds to the standard interference limited regime with no interference management, while the second case is an ideal case with full interference removal [13]. The main important figure of merit is that in the second case, the same fair capacity may be achieved with a 30 dB reduction of total transmitted power. When using efficient resource sharing techniques exploiting either orthogonal frequency-division multiple access (OFDMA) or superposition coding (SC), the theoretical EE gain is huge: in full interference, EE reaches a maximum of 60 Mb/J while it grows to 30 Gb/J with no interference, as seen in Fig. 4. (Note here that the EE metric includes only the transmit power. If the total energy were considered, EE would decrease by at least a factor of 10.)

However, achieving full interference cancellation is infeasible in practice. Interference is produced by a large set of neighbor BSs, including a few strong interferers (the nearest BS) and a bunch of long distance interferers. However, it is estimated that removing the two or three strongest interferers may still provide a gain of about 15 dB [13]. To achieve this gain, different interference management schemes may be used and combined, such as resource partitioning, scheduling, beamforming, cooperative transmission, and interference alignment. These schemes have different coordination requirements and effective gain. In 3GPP, enhanced intercell interference coordination (eICIC) techniques are introduced, which only require minimum coordination between cells.

When full feedback is available, the multi-cell scenario becomes equivalent to a single-cell MIMO scenario, and the optimal performance may be achieved by employing joint transmission and reception of multiple cells. To reduce the amount of backhaul transmissions and cooperation requirements, interference alignment is attractive because it requires only information exchange relative to channel states, each mobile being associated with only one BS. However, the performance of interference alignment is limited by imperfect channel estimation and time variations. Interference alignment can also be designed with limited feedback, but at the price of reduced DoFs. Indeed, the local choice made by a transmitter should be restricted to not affect the interference perceived by neighbor receivers.

There are other effective intercell interference management schemes, but there is no room to enumerate all of them. Future work on intercell interference may include:

EE-SE analysis: Theoretical analysis of EE-SE trade-off in a multi-cell scenario will be valuable and interesting work, which can provide insights on the performance bounds and which schemes are more promising from the viewpoint of EE-SE trade-off. The difficulty mainly lies in how to balance the complexity and model accuracy.

Intercell coordination: The coordination between cells is the major limitation of intercell interference management schemes. The multi-cell network is a dynamic system with local agents. The appropriate decision loops, feedback, and decision rules will play a critical role. Furthermore, the delay and limited capacity of backhaul connections have to be considered.

Distributed Antenna Systems

Distributed antenna systems deploy antennas in a distributed manner, so that the network coverage is increased. Since the antennas are closer to the

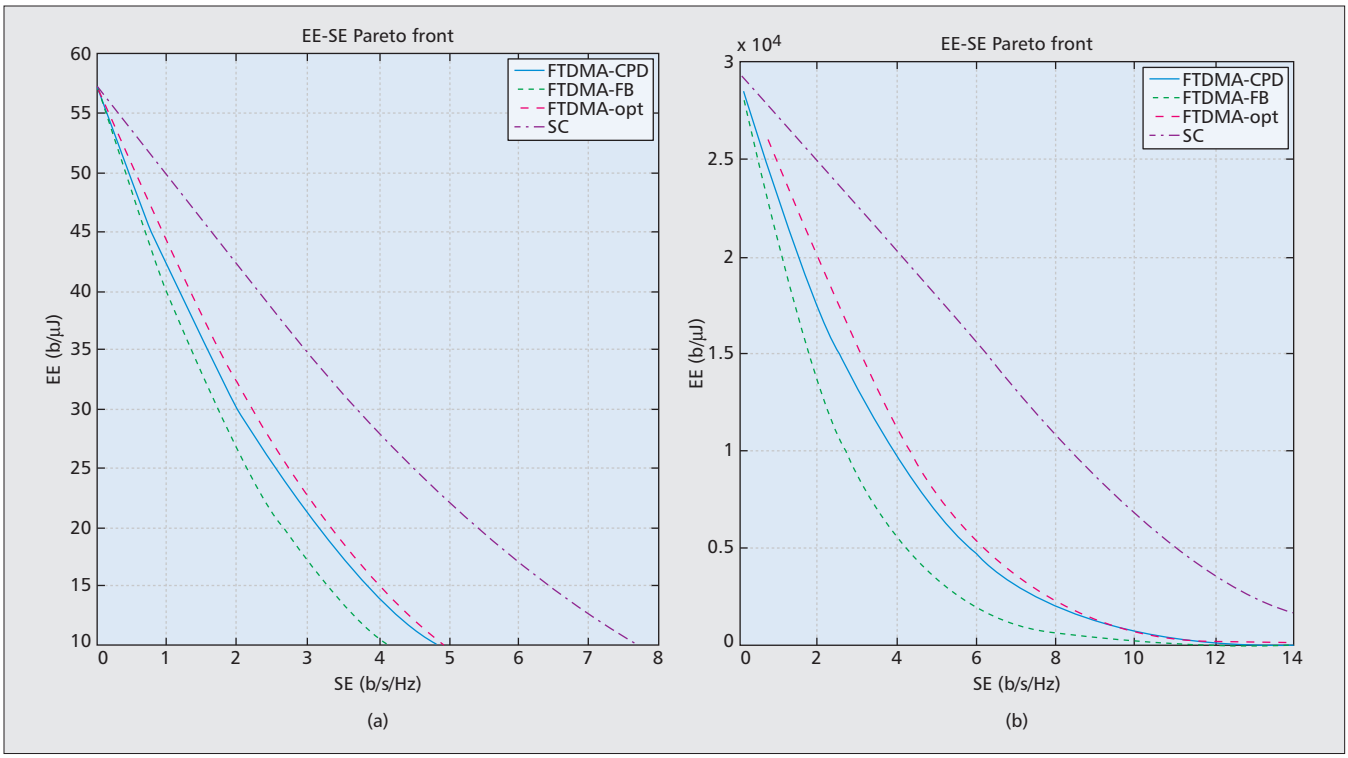


Figure 4. EE-SE trade-off with different resource sharing techniques (FTDMA-CPD, -FB, and -opt: time-frequency division, with constant power density, constant bandwidth per user, and optimal allocation, respectively. SC : superposition coding): a) full reuse scenario; b) interference-free scenario.

users, the path loss between transmitters and receivers decreases. Under the EE-SE trade-off framework, both EE and SE can be improved with DASs. Furthermore, a DAS has a novel structure with central baseband units (BBUs) and remote radio units (RRUs), which largely reduces the number of BS sites while maintaining site capacity. As a result, network power consumption of supporting equipment such as cooling can be largely reduced. Due to the limitation of computation power, each BBU can only support several RRUs. Large-scale DASs will become feasible with more and more powerful processors. Figure 5a shows an implementation architecture of large-scale DAS, in which a number of RRUs are connected to a pool of BBUs via optical switches. These high-throughput optical switches can dynamically forward and receive real-time data of each RRU to/from its related BBU.

Intuitively, the best performance can be achieved in a DAS by coordinating all the RRUs together and performing distributed beamforming. However, this type of coordination can induce considerable computational complexity. To reduce the complexity, antenna clustering according to user location and/or channel quality can be applied. With antenna clustering, intra-cluster interference can be eliminated with the cooperative transmission schemes introduced previously, but inter-cluster interference still exists since there is no coordination among the clusters. As the BBUs are co-located, a central controller can be applied to achieve inter-cluster coordination so that the interference can be further reduced.

Our recent work in [14] proposed a resource allocation scheme for large-scale DASs, in which hundreds of antennas are randomly deployed, which jointly considers the power allocation and preceding MIMO. A simplified cloud-RAN-based energy-efficient power allocation (S-CEEPA) scheme is proposed in which instantaneous CSI is not required. Thus, this scheme can easily be employed in frequency-division duplex (FDD) mode. The uncontrollable inter-cluster interference is also taken into account. The baseline scheme is equal power allocation (EPA) for each antenna. The simulation results in Fig. 5b show that the proposed scheme performs much better than the baseline scheme. For all the scenarios, EE increases with more antennas deployed. This is because with increasing antenna density, the propagation loss decreases, and the interference is largely eliminated by intra-cluster precoding. On the other hand, EE decreases when the number of clusters increases due to higher inter-cluster interference. Future work on distributed antennas systems includes:

Power model: Since the network architecture is different from traditional wireless networks, the power model of large-scale DASs needs to be further investigated. On one hand, the power consumption of support equipment can be reduced with the structure of centralized processing. On the other hand, additional power consumption is caused by optical switches, BBUs, and the controller for intra-cluster and inter-cluster signal processing.

Imperfect transmission: The RRUs are assumed to be perfectly synchronized, and the overhead of CSI acquisition is neglected in current work. If the transmissions are not perfect,

The GTT toolbox provides a simulation framework to evaluate the performance of integrated GTT solution, which is affected by variable factors, such as network scenarios, traffic dynamics, large-scale and small-scale fading. There is a graphic user interface to configure all these factors.

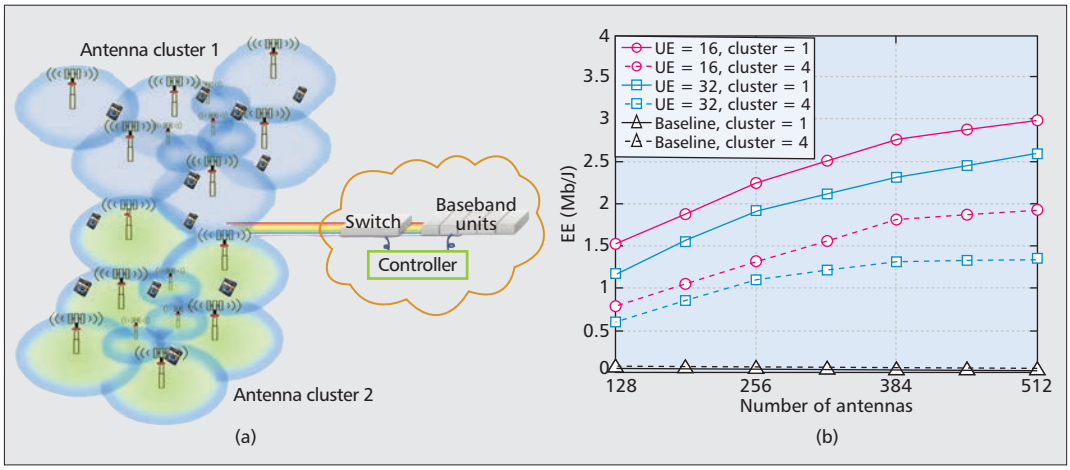


Figure 5. a) Architecture of a large-scale distributed antenna system; b) energy efficiency with different numbers of antennas.

the intra-cluster interference cannot be fully eliminated. To investigate the practical benefit of large-scale DASs, it is necessary to study the optimization of joint clustering and precoding design under imperfect transmission.

Integration of GTT Solutions

In the previous section, various GTT solutions have been proposed to improve EE-SE trade-off for wireless networks. For each scenario, only one or several solutions can be used. One way to evaluate the EE-SE trade-off with the combined effects of all GTT solutions is to develop a system-level simulator with all the solutions applied. However, such a simulator is too complicated and time-consuming. In this section, the GTT toolbox is introduced as a systematic simulation tool and unified simulation platform to integrate all these GTT solutions together, as well as the method of integration.

GTT Toolbox

The GTT toolbox provides a simulation framework to evaluate the performance of an integrated GTT solution, which is affected by variable factors, such as network scenarios, traffic dynamics, and large-scale and small-scale fading. There is a graphic user interface to configure all these factors, as shown in Fig. 6a. The network scenario and traffic model are both configurable, and the channel models include propagation loss, large-scale fading, and small-scale fading, which mainly follow the parameters specified in [15]. The power model is either a linear model or other more sophisticated models with configurable parameters. If the "Go" button is pressed, a system-level simulation is executed. After the simulation, different performance metrics, including EE, SE, user throughput, and power consumption, are illustrated.

Method of Integration

The main difficulty of system-level simulation lies in the time-varying states. If traffic dynamics is considered, the dimension of states becomes intractable. To solve this problem, we propose a new simulator structure in the GTT toolbox, as illustrated in Fig. 6b. The simulator consists of the network and physical layers. Instead of implementing the GTT solutions in detail, they are abstracted into signal-to-interference-plus-noise ratio (SINR)-SE or signal-to-noise ratio (SNR)-interference-to-noise ratio (INR)-SE mapping curves or look-up tables in the physical layer. Each GTT solution may have multiple mapping curves, and each mapping curve corresponds to a set of parameters (e.g., the number of transmit and receive antennas).

These mapping curves act as an interface between the network and physical layers. In the network layer simulation, traffic dynamics and large-scale fading are considered. When a user arrives, the SINR, SNR, and INR are calculated based on the large-scale fading model. The average SE is given by the mapping curves, which are functions of SINR, SNR, and INR. Usually, the curve with the best SE is selected. If multiple users are served by the same cell, the bandwidth is equally allocated to the users. Thanks to the two-layer structure, the scheduling or bandwidth allocation issues are left to the physical layer simulation.

The mapping curves are obtained by offline simulation, in which the number of users is fixed and all the cells are under full load. Both large-scale and small-scale fading are considered. At the beginning of simulation, a number of users are randomly dropped, and the average SINR, SNR, and INR of each user is calculated based on the large-scale fading model, which corresponds to the horizontal axis of the mapping curves. Then the users are served by the BSs, and the instantaneous SE for each scheduled user is stored. The pairs of average SE and average SINR form scatter plots, and the mapping curves are computed by curve fitting at the end of the simulation.

If MU-MIMO is applied, users can have more transmission opportunities by bandwidth sharing. However, this is not reflected in the calculation of average SE. The instantaneous SE is multiplied by the number of users if MU-MIMO is applied to reflect multiuser diversity. Fig. 6c illustrates an example of multiple SINR-SE curves for MU-MIMO. As MU-MIMO can only be applied if there are multiple active users, thus

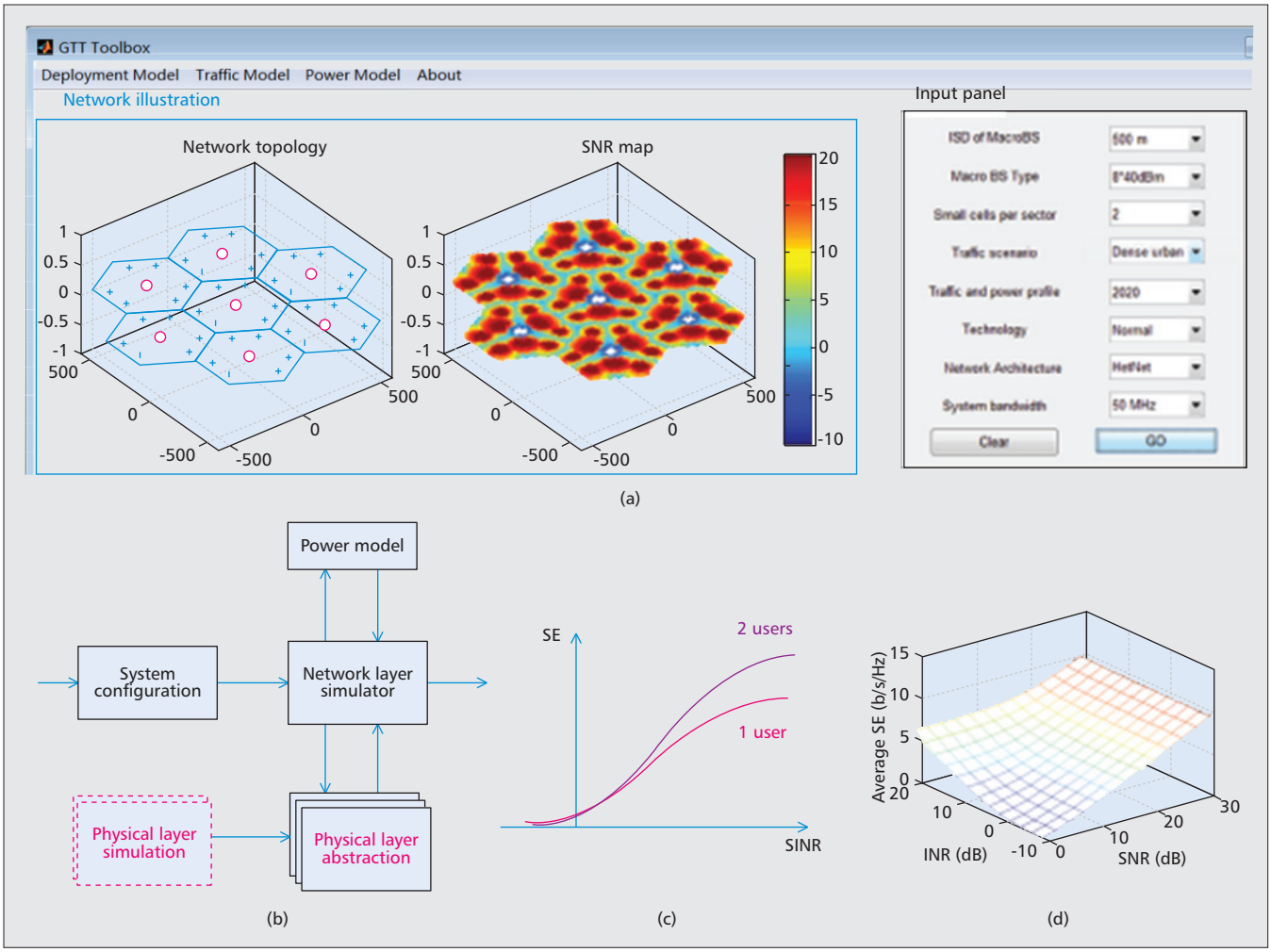


Figure 6. a) User interface of the GTT toolbox; b) simulator structure of the GTT toolbox; c) the example of SINR-SE curves for MU-MIMO; d) the example of SNR-INR-SE curves for multi-cell solutions.

for network-layer simulation, the mapping curves are not applicable if there are not enough users.

If intercell interference management is applied, SINR can be improved, especially for cell edge users. Since the average SINR is calculated with all interference, users who suffer from high interference and under deep fade cannot be distinguished with the SINR-SE curve. Therefore, we develop a new kind of mapping curve, the SNR-INR-SE curve, shown in Fig. 6d. In this way, the abstraction is more accurate for multi-cell GTT solutions.

It should be noted that the above method is only an approximation of the real system-level simulation. The approximation holds only if the timescale of network-layer dynamics is much longer than the physical layer (i.e., the service rate of each user can be approximated by the average performance. As a whole, the GTT toolbox provides a simple and flexible way to integrate multiple GTT solutions, which play an important role in the whole GTT project.

Conclusions

This article has described the design principle of energy-efficient wireless networks in the GTT project. Inspired by the fundamental framework of EE-SE trade-off, four selected GTT solutions have been introduced, including bandwidth expansion, multiple-antenna technologies, intercell interference management, and distributed antenna systems. The performance benefits, as well as design insights and challenges of these solutions have also been elaborated. Furthermore, the GTT toolbox and a novel method of integration have been introduced to integrate all the GTT solutions. Future work in the GTT project will focus on the design of integrated green transmission technologies and provision of a systematic solution for the big challenge of energy consumption in current and future 5G wireless networks.

References

[1] E. Oh *et al.*, "Toward Dynamic Energy-Efficient Operation of Cellular Network Infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 56–61.
[2] Huawei, "5G: A Technology Vision," http://www.huawei.com/5gwhitepaper, 2013.
[3] C. Han *et al.*, "Green Radio: Radio Techniques to Enable Energy-Efficient Wireless Networks," *IEEE Commun. Mag.*, vol. 49, no. 6, Jun. 2011, pp. 46–54.
[4] Y. G. Li *et al.*, "Energy-Efficient Wireless Communications: Tutorial, Survey, and Open Issues," *IEEE Wireless Commun.*, vol. 18, no. 6, Dec. 2011, pp. 28–35.
[5] Z. Niu *et al.*, "Cell Zooming for Cost-Efficient Green Cellular Networks," *IEEE Commun. Mag.*, vol. 48, no. 11, Nov. 2010, pp. 74–79.

Future work of the GTT project will focus on the design of integrated green transmission technologies and provision of a systematic solution for the big challenge of energy consumption in current and future 5G wireless networks.

[6] L. M. Correia *et al.*, "Challenges and Enabling Technologies for Energy Aware Mobile Radio Networks," *IEEE Commun. Mag.*, vol. 48, no. 11, Nov. 2010, pp. 66–72.
[7] T. Chen *et al.*, "Network Energy Saving Technologies for Green Wireless Access Networks," *IEEE Wireless Commun.*, vol. 18. no. 5, Oct. 2011, pp. 30–38.
[8] C. L. I *et al.*, "Toward Green and Soft: A 5G Perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 66–73.
[9] Y. Chen *et al.*, "Fundamental Trade-offs on Green Wireless Networks," *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 30–37.
[10] G. Auer *et al.*, "How Much Energy is Needed to Run A Wireless Network?" *IEEE Wireless Commun.*, vol. 18. no. 5, Oct. 2011, pp. 40–49.
[11] J. Tang *et al.*, "Resource Efficiency: A New Paradigm on Energy Efficiency and Spectral Efficiency Tradeoff," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, Aug. 2014, pp. 4656–69.
[12] M. Butt *et al.*, "On the Energy-Bandwidth Tradeoff in Green Wireless Networks: System Level Results," *Proc. IEEE ICCC Wksp.*, Aug. 2012, pp. 91–95.
[13] J. M. Gorce *et al.*, "Energy-Capacity Trade-Off Bounds in a Downlink Typical Cell," *Proc. IEEE PIMRC*, Sept. 2014.
[14] P. R. Li, T. S. Chang, and K. T. Feng, "Energy-Efficient Power Allocation for Distributed Large-Scale MIMO Cloud Radio Access Networks," *Proc. IEEE WCNC*, Apr. 2014.
[15] 3GPP TR 36.814, "Evolved Universal Terrestrial Radio Access (E-UTRA): Further Advancements for E-UTRA Physical Layer Aspects," Mar. 2010.

Biographies

Yiqun Wu (wuyiqun@huawei.com) received B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, China, in 2006 and 2012, respectively. He was a visiting researcher at the Chinese University of Hong Kong from 2007 to 2008 and at Ohio State University from 2010 to 2011. Since 2012, he has been with Huawei Technologies Co., Ltd. Shanghai. His research interests include energy-efficient wireless networks, new waveforms, and multiple access schemes for 5G.

Yan Chen received her B.Sc. and Ph.D. degrees in information and communication engineering from Zhejiang University, China, in 2004 and 2009, respectively. She was a visiting researcher at Hong Kong University of Science and Technology (HKUST), from 2008 to 2009. She joined Huawei Technologies Co., Ltd. Shanghai in 2009 and has been the team leader for green radio research since then. She is now working as the project manager on green air interface design for fifth generation mobile communications in Huawei and project leader for the GTT project in GreenTouch.

Jie Tang [S'10, CM'13] is a postdoctoral research associate in the School of Electrical & Electronics Engineering, University of Manchester, United Kingdom. He received his Ph.D. degree from Loughborough University, United Kingdom, his M.Sc. degree from the University of Bristol, United Kingdom, and his B.Eng. degree from the South China University of Technology, China. His current research interests include green communications, 5G systems, heterogeneous networks, cognitive radio, and MIMO systems.

Daniel K. C. So [S'96, M'03, SM'14] is a senior lecturer in the School of Electrical and Electronic Engineering, University of Manchester. He received his Ph.D. degree from HKUST, and his B.Eng. degree from the University of Auckland, New Zealand. His current research interests include green communications, 5G networks, heterogeneous networks, massive MIMO, and cognitive radio.

Zhikun Xu received his B.S.E. and Ph.D. degrees in electrical and computer engineering from Beihang University, China, in 2007 and 2013, respectively. He was a visiting researcher in the School of Electrical and Computer Engineering, Georgia Institute of Technology, from 2009 to 2010. After graduation, he joined the Green Communication Research Center of the China Mobile Research Institute as a project manager. His current research mainly focuses on 5G technologies, including energy efficiency and spectral efficiency co-design, large-scale antenna systems, non-orthogonal multiple access schemes, and cross-layer resource allocation.

Chih-Lin I received her Ph.D. degree in electrical engineering from Stanford University and has almost 30 years of experience in wireless communications. She has worked at various world-class companies and research institutes, including the Wireless Communication Fundamental Research Department of AT&T Bell Labs, the headquarters of AT&T, ITRI of Taiwan, and Hong Kong ASTRI. She received the *IEEE Transactions on Communications* Stephen Rice Best Paper Award and is a winner of the CCCP National 1000 Talent program. Currently, she is China Mobile's chief scientist of wireless technologies in charge of advanced wireless communication R&D efforts of the China Mobile Research Institute. She established the Green Communications Research Center of China Mobile, spearheading major initiatives including 5G key technologies R&D; high energy efficiency system architecture, technologies, and devices; green energy; and C-RAN and soft base stations. She was an elected Board Member of IEEE ComSoc, Chair of the ComSoc Meetings and Conferences Board, and Founding Chair of the IEEE WCNC Steering Committee. She is currently an Executive Board Member of GreenTouch and a Network Operator Council Member of ETSI NFV. Her research interests are green communications, C-RAN, network convergence, bandwidth reframing, EE-SE co-design, massive MIMO, and active antenna arrays.

Jean-Marie Gorce received M.S. and Ph.D. degrees in electrical engineering from the National Institute of Applied Sciences (INSA), Lyon, France, in 1993 and 1998. After a postdoctoral year at Bracco Research, Switzerland, he joined the Telecommunications Department at INSA Lyon as an associate professor, where he is head of the radio modeling axis of CITI Laboratory. His main research field concerns wireless networks focusing on realistic modeling, wireless system optimization, and performance assessment, considering as well architecture-based and ad hoc networks.

Chih-Hsuan Tang received her Ph.D. degree from National Chiao Tung University, Hsinchu, Taiwan, in 2007. Since 2007, she has been a researcher at Chunghwa Telecom Laboratories, Taiwan. Her current research interests include green communications, self-optimization networks, and heterogeneous networks.

Pei-Rong Li received her B.S. degree from Yuan Ze University, Taoyuan, Taiwan, in 2012. Since 2014, she has been working toward a Ph.D. degree in the Department of Communications Engineering, National Chiao Tung University. Her research interests include resource allocation and spectrum sharing for cooperative networks.

Kai-Ten Feng received his M.S. degree from the University of Michigan, Ann Arbor, in 1996, and his Ph.D. degree from the University of California, Berkeley, in 2000.Since August 2011, he has been a full professor with the Department of Electrical and Computer Engineering, National Chiao Tung University. His current research interests include broadband wireless networks, cooperative and cognitive networks, and wireless location technologies.

Li-Chun Wang [M'96, SM'06, F'11] received his Ph. D. degree from the Georgia Institute of Technology in 1996. He was with AT&T Laboratories from 1996 to 2000, and is the current Chairman of the Department of Electrical and Computer Engineering of National Chiao Tung University. His current research interests are in the areas of radio resource management and cross-layer optimization techniques for heterogeneous wireless networks, and cloud computing for mobile applications.

Kai Börner received his M.S. degree in electrical engineering from Technische Universität Berlin, Germany, in 2009. Currently, he is working toward a Ph.D. degree in electrical engineering at Technische Universität Berlin. He joined Fraunhofer Heinrich Hertz Institute (HHI) in April 2009. His research interests lie in channel modeling, energy-efficient transmission, and self-organization in heterogeneous MIMO-OFDM-based networks.

Lars Thiele received his M.S. degree in electrical engineering from Technische Universität Berlin in 2005. He joined Fraunhofer HHI in September 2005. In 2013 he received his Ph.D. degree from the Technical University of Munich. He has contributed to receiver and transmitter optimization under limited feedback, performance analysis for MIMO transmission in cellular OFDM systems, fair-resource allocation, and cooperative multi-point transmission under constrained channel state information at the transmitter.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
COMMUNICATIONS STANDARDS SUPPLEMENT

BACKGROUND

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

SCOPE OF CONTRIBUTIONS

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

Analysis of new areas for standardization, either enhancements to existing standards or new areas. 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 is not limited to:
- The national, regional, and global impacts of standards on industry, society, and economies
- The processes and organizations for creation and diffusion of standards, including the roles of organizations such as IEEE and IEEE-SA
- National and international policies and regulation for standards
- Standards and developing countries

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

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

Compatibility and interoperability, including testing methodologies and certification to standards

Tools and services related to any or all aspects of the standardization 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 dropdown menu of submission options.

A Survey of Energy-Efficient Caching in Information-Centric Networking

Chao Fang, F. Richard Yu, Tao Huang, Jiang Liu, and Yunjie Liu

ABSTRACT

To better cope with the Internet usage shift from host-centric end-to-end communication to receiver-driven content retrieval, innovative information-centric networking (ICN) architectures have been proposed. A notable advantage of these novel networking architectures is to provide transparent and ubiquitous in-network caching to speed up content distribution and improve network resource utilization. With the explosive increase of global network traffic, the energy efficiency issue in ICN is a growing concern. In this article, we provide a brief survey of energy-efficient caching techniques in ICN from the placement, content placement, and request-to-cache routing perspectives. We also outline some challenges and future research directions about caching policies for green ICN.

INTRODUCTION

According to the Cisco Visual Networking Index 2013, global IP traffic has increased more than fourfold in the past five years, and will grow threefold by 2017, to reach 120.6 exabytes per month. Most of the traffic can be attributed to the emergence of video streaming portals for user generated content (e.g., YouTube and Google Video) and video on demand (e.g., NetFlix, Hulu, and IPTV services). It is predicted that global consumer Internet video traffic will be 69 percent of all consumer Internet traffic in 2017, up from 57 percent in 2012, and the sum of all forms of video (e.g., TV, VoD, Internet, and P2P) will be in the range of 80–90 percent of global IP traffic by 2017. While IP has exceeded all expectations for facilitating ubiquitous interconnectivity, it was designed for conversations between communications endpoints, but is overwhelmingly used for content distribution. That is, today's Internet is increasingly used for information dissemination rather than pair-wise communications between end hosts [1], and pays more attention to the content itself rather than where it is physically located [2]. However, the current Internet, originally conceived to enable communication between machines, lacks natural support for content distribution. This fundamental mismatch has significant impacts on network performance in terms of end-user quality of experience, bandwidth costs, delay, and energy use [3].

Initial attempts to accommodate content distribution within the Internet infrastructure have resulted in a plethora of content-oriented applications/services, such as peer-to-peer (P2P) networks (e.g., Gnutella and BitTorrent) and content delivery networks (CDNs). In these application-/service-specific solutions, an end user does not care about hosts, but about content. Therefore, these mechanisms can improve content access and quality of user experience over the Internet. However, from a networking perspective, they still rely on a host-to-host communication model and do not take into account social semantics of transferred content for optimizing content routing or caching, which leads to costly and/or inefficient solutions for content distribution [4].

To better cope with the Internet usage shift from a sender-driven end-to-end communication paradigm to a receiver-driven content retrieval one, a handful of innovative information-centric networking (ICN) architectures have been proposed [1]. The philosophy behind ICN is to promote content to a first-class citizen in the network. In ICN, users do not care where the content comes from, but are only interested in what the content is. The essence of ICN lies in decoupling contents from hosts (or their locations) not at the application level, but at the network level. A notable advantage of these novel networking architectures is to provide transparent and ubiquitous in-network *caching* to speed up content distribution and improve network resource utilization, as requests no longer need to travel to the content source, but are typically served by a closer ICN content router along the routing path [5].

Although some excellent works have been done on ICN in-network caching, their focus is to increase cache hit rate and/or decrease network delay. Consequently, the energy consumption aspect in this setting is largely ignored [6]. However, the increasingly strict environmental standards and rapidly rising energy costs have led to an emerging trend of addressing the energy efficiency of the Internet. The information and communication technologies (ICT) sector is

Chao Fang, Tao Huang, Jiang Liu, and Yunjie Liu ,are with Beijing University of Posts and Telecom.

F. Richard Yu is with Carleton University.

0163-6804/14/$25.00 © 2014 IEEE

currently responsible for almost 2 percent of world electricity use in 2007, having observed an annual increase of 10 percent from 2007 to 2012. The total global electricity and diesel energy consumption by all mobile networks was approximately 120 TWh in 2010, resulting in energy costs of US$13 billion and responsible for 70 Mt carbon dioxide equivalent (CO_{2e}). In a business as usual (BAU) scenario, the non-negligible greenhouse gas (GHG) emissions of the ICT sector are expected to reach 1.43 Gtons CO_{2e} in 2020.

In this article, we provide a brief survey of energy-efficient caching techniques in ICN. The rest of this article is organized as follows. We first present the history and main components of ICN, and then introduce the architecture and workflow of the typical content-centric networking-based ICN approach. After introducing the ICN paradigm, we describe ICN in-network caching. We then present the existing energy-efficient techniques from the cache placement, content placement, and request-to-cache routing perspectives. We also outline some research challenges and future research directions for energy-efficient ICN caching. Finally, we conclude the article.

An Overview of Information-Centric Networking

This section introduces the history and main components of ICN, and presents the architecture and workflow of the typical content-centric networking-based ICN approach.

A Brief History of Information-Centric Networking

ICN has been attracting increasing attention from both academia and industry. A number of ICN initiatives focus on designing an Internet architecture that can replace the current host-centric model and directly address the content delivery problem described above. The concept of ICN dates back to 1999, when Cheriton *et al.* introduced the concept of name-based routing in translating relaying Internet architecture integrating active directories (TRIAD), which proposed to avoid DNS lookups by using the name of an object to route toward a close replica of it. In 2007, data-oriented network architecture (DONA) was proposed as one of the first clean-slate ICN proposals. DONA uses flat, self-identifying, and unique names for information objects, and binds the act of resolving requests for information to locating and retrieving information, which improves TRIAD by incorporating security (authenticity) and persistence as first-class primitives in the architecture.

Subsequently, a number of research efforts have been dedicated to ICN, including the EU funded projects Publish-Subscribe Internet Technology (PURSUIT) and its predecessor Publish-Subscribe Internet Routing Paradigm (PSIRP), Scalable & Adaptive Internet Solutions (SAIL) and its predecessor 4WARD, Content Mediator Architecture for Content-Aware Networks (COMET), CONVERGENCE, the U.S. funded projects Named Data Networking (NDN) and its predecessor Content-Centric Networking (CCN) and MobilityFirst, the French funded project ANR Connect, which adopts the NDN architecture, as well as the collaborative EU-Japan project called GreenICN.

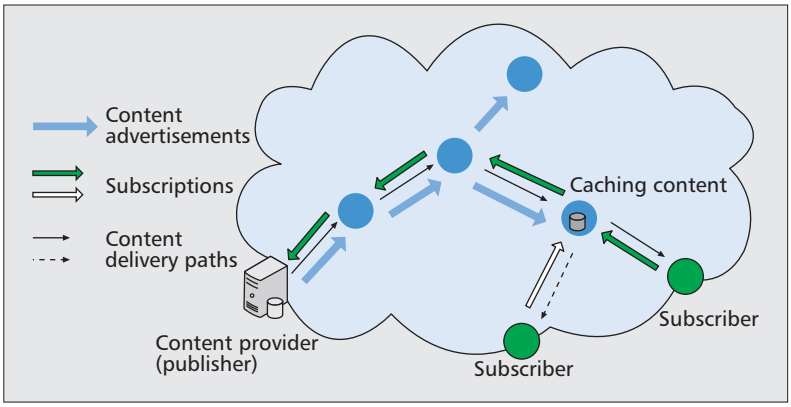


Figure 1. An information-centric network.

The Main Components of Information-Centric Networking

The communication paradigm within ICN is different from that of IP. Current IP architectures revolve around a host-based conversation model (i.e., a communication is established between two hosts before any content is transferred), and the delivery of data in the network follows a source-driven approach (i.e., the path is set up from the sender to the receiver). The principal concern of ICN is to disseminate, find, and deliver information rather than the reachability of end hosts and the maintenance of conversations between them. In ICN, the user requests content without knowledge of the host that can provide it, the communication follows a receiver-driven principle (i.e., the path is set up by the receiver to the provider), and the data follows the reverse path. The network is then in charge of doing the mapping between the requested content and where it can be found (Fig. 1). The match of requested content rather than the findability of the endpoint that provides it thus dictates the establishment of a communication in ICN.

To be efficient, one key aspect of ICN is naming. Content should be named in such a way as to be independent of the location of the node where the content can be found, which is the main objective of ICN (to separate naming and location). ICN also includes a native caching function in the network in such a way that nodes can cache the contents passing through it for a while (depending on the cache size and replacement algorithm) and deliver them to requesting users. Via this in-network caching mechanism, the content is replicated, and the delivery probability of the content to the end user is increased.

Decoupling naming from location also allows native support of mobility or multicast in ICN. Indeed, when users move, they are connected to another node in the ICN network, but since no IP address is used for the routing, it is transpar-

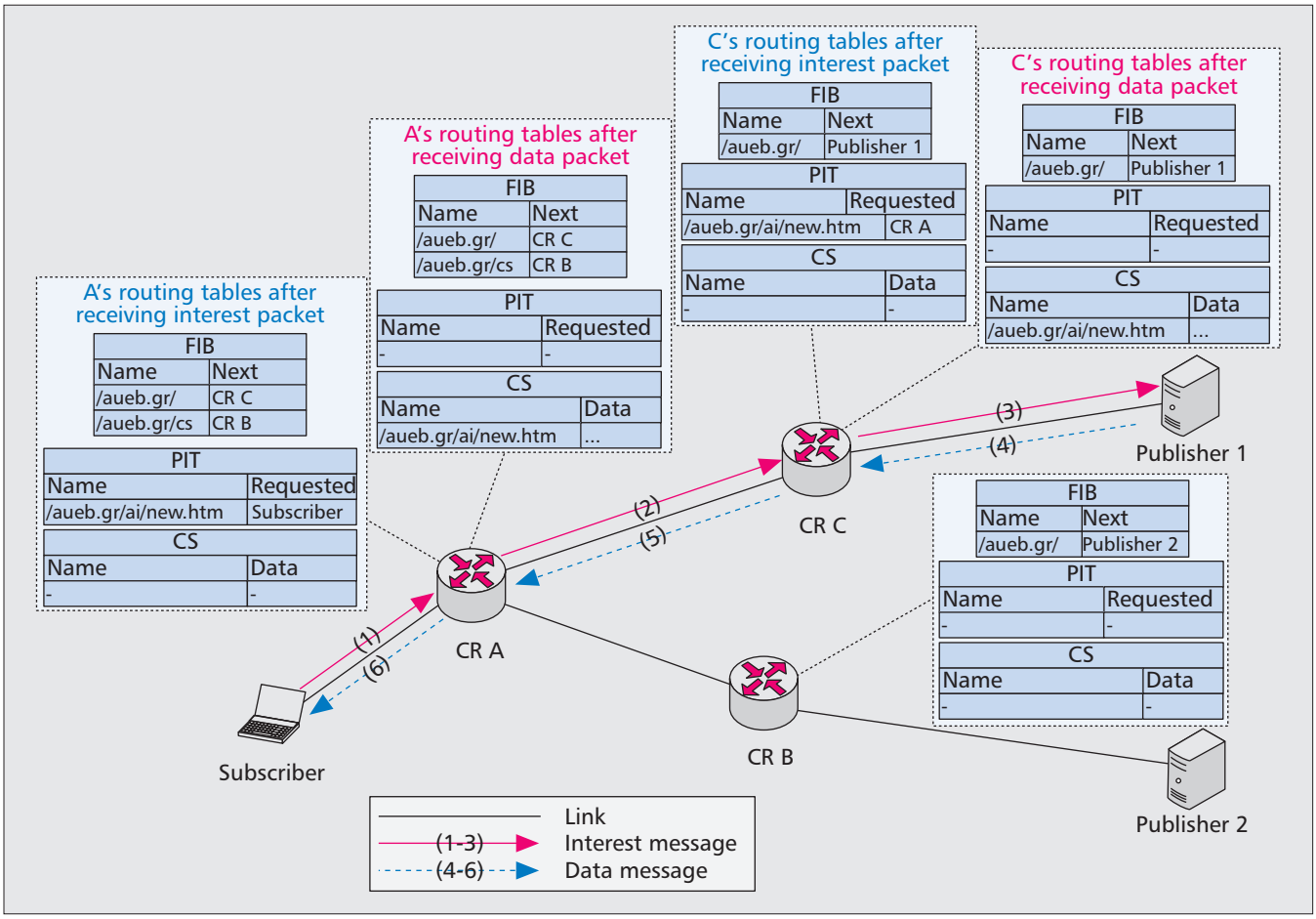


Figure 2. Content-centric networking-based ICN architecture. CR: content router; FIB: forwarding information base; PIT: pending interest table; CS: content store.

ent, as opposed to IP, where the address should be changed. For multicast, as soon as one user has requested a given content, one node can cache it and then deliver it for subsequent requests for the same content. It then naturally creates multicast-like content delivery.

Introduction to Content-Centric Networking: A Pioneer Information-Centric Networking Solution

The open source implementation of CCN describes a complete naming scheme, content retrieval storage, and dissemination algorithms, and this implementation is also the root of the NDN project proposing a comprehensive networking protocol designed around CCN. Consequently, CCN has become one of the most promising techniques for ICN [4, 7]. Therefore, in this subsection, we present the architecture and workflow of CCN-based ICN approach.

Architecture of Content-Centric Networking — The CCN-based ICN architecture is showed in Fig. 2. CCN is a receiver-driven data-centric communication protocol. Communication in CCN is performed using two distinct types of packets: *interest packets* and *data packets*. Both types of packets carry a name, which uniquely identifies a piece of data that can be carried in one data packet. Besides, to receive data, each CCN content router (CR) maintains three major data structures: a content store (CS) for temporary caching of received data packets, a pending interest table (PIT) to contain the names of interest packets and a set of interfaces from which the matching interest packets have been received, and a forwarding information base (FIB) to forward interest packets.

The Workflow of Content-Centric Networking — Figure 2 also shows the working procedure of the CCN-based ICN architecture. The subscriber sends an Interest for the name /aueb.gr/ai/new.htm (arrows 1–3). When the interest packet arrives, the CR extracts the information name and looks for an information object in its CS with a name matching the requested prefix. If something is found, it is immediately sent back through the incoming interface in a data message, and the interest packet is discarded. Otherwise, the router performs a longest prefix match on its FIB in order to decide in which direction this interest packet should be forwarded. If an entry is found in the FIB, the router records the interest packet's incoming interface in the PIT and pushes the packet to the CR indicated by the FIB.

When an information object that matches the requested name is found at a publisher node or

a CS, the Interest message is discarded and the information is returned in a Data message. When a CR receives a Data message, it first stores the corresponding information object in its CS and then performs a longest-prefix match in its PIT to locate an entry matching the data packet; if a PIT entry lists multiple interfaces, the Data message is duplicated, thus achieving multicast delivery. Finally, the CR forwards the Data message to these interfaces and deletes the entry from the PIT (arrows 4–6). If there are no matching entries in the PIT, the router discards the data packet as a duplicate.

In-Network Caching of Information-Centric Networking

In this section, we first introduce the advantages of in-network caching in ICN, and then present two approaches to in-network caching.

Advantages of In-Network Caching

In-network caching is a fundamental feature of ICN architectures, as information awareness allows the network to identify cached information without resorting to the application layer, as in web caching. Therefore, it can improve network performance by fetching content from nodes geographically placed closer to the end user. An illustration of content caching in CCN is shown in Fig. 3. The usefulness of caching is already proven by the commercial success of CDNs. ICN generally leverages in-network storage to provide a better-performing and more robust transport service. For example, the advantages of in-network caching for an Internet service provider (ISP) may be twofold: reducing the incoming traffic from neighbor ISPs to lower the traffic load on its cross-ISP links (and hence its expense for transport link capacity) and improving the delay/throughput performance by placing the contents closer to their users. In-network caching is also attractive to content providers (CPs) since it can mitigate the capital expense of their content servers.

On-Path and Off-Path Caching

There are two approaches to in-network caching: on-path and off-path caching. On-path caching is generally opportunistic (i.e., routers cache information that happens to flow through them), while off-path caching can be used to actively replicate information, as in CDNs.

On-Path Caching — A straightforward approach to content placement is on-path placement of contents as they travel from source to destination. In on-path caching, when a router receives a request for a piece of information, it responds with a locally cached copy without involving the name resolution system. Although this approach reduces the computation and communication overhead of placing content within the network, it might reduce the chances of hitting cached contents.

An important issue in on-path caching is how each node makes a caching decision to improve the cache hit rate of content delivery. For example, popular content might need to be placed where it is going to be requested next. Furthermore, problems of expected content popularity or temporal locality need to be taken into account in designing in-network caching algorithms in order for some contents to be given priority (e.g., popular contents vs. one-timers). The criteria as to which contents should be given priority in in-network content caches also relate to the business relationships between content providers and network operators.

While all ICN architectures natively support on-path caching in principle, when name resolution and data routing are decoupled there are fewer opportunities to exploit opportunistic caching, as the name resolution path generally differs from the data routing path: while the information can be opportunistically cached on the data routing path, subsequent requests for the same information follow the (different) name resolution path, reducing the possibility for a cache hit. However, when name resolution and data routing are coupled, if data is cached on the data routing path it will result in a cache hit when subsequently requested over the same name resolution path. Opportunistic caching can range from the "cache everything" approach of CCN to the probabilistic caching approach of COMET [8].

Off-Path Caching — Off-path caching is similar to traditional proxy caching or CDN server placement. In off-path caching, caches announce

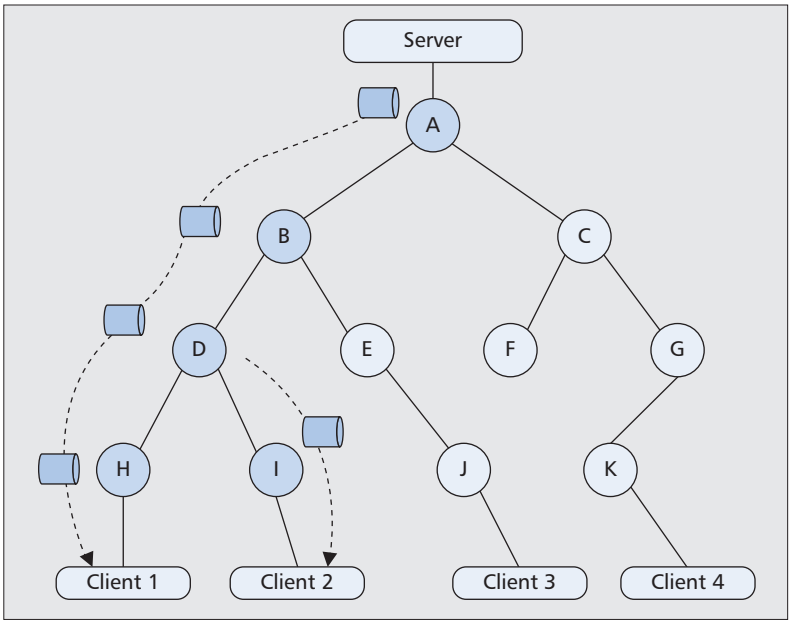


Figure 3. Illustration of content caching in CCN: Client 1 fetches a content by sending an interest packet. Initially, were no other interest packets were issued earlier, and intermediate ICN routers do not have the requested chunk in their buffers. The interest packet is eventually delivered to the origin server along the shortest path (i.e., H, D, B, and A). The requested chunk is then delivered by traversing the reverse path (i.e., A, B, D, and H), and each intermediate router keeps the forwarded chunk in its router buffer. If client 2 wants to access the same content, its interest packet will find a match at node D, and the requested chunk will be delivered directly from that node.

their information to the name resolution system so that they may be matched to information requests that would not normally reach them, essentially becoming alternative information publishers. Therefore, retrieval of contents from off-path caches requires redirection of requests.

Beyond the more general problem of choosing what to cache and where, the main issue in off-path caching is how to reduce the overhead required in order to inform the name resolution system when new items are cached or old items are discarded. The exact details depend on the name resolution scheme used, but one common goal is to keep updates local, for example, within an autonomous systems (AS) in order to reduce signaling overhead and only serve customers from within that AS [8]. In DONA and COMET, cached information can be advertised only within an AS and not propagated upward in the AS hierarchy (COMET provides the scope mechanism for this purpose). Similarly, in PURSUIT and the decoupled version of SAIL, cached information can only be advertised within the local distributed hash table (DHT) of an AS. In CCN, CONVERGENCE, and the coupled version of SAIL, the name prefix tables need to be updated, but it is unclear how this could be achieved economically as the routing protocols proposed for advertising name prefixes are based on flooding. MobilityFirst also faces problems in this area, as it relies on a global lookup mechanism for name resolution; therefore, it is unclear how locally cached copies can be advertised only within an AS.

Technology	Reference	Contributions
Energy-efficient cache placement	Braun *et al.* [6]	Analyzing the access frequency threshold from which caching content starts to be beneficial.
	Chen *et al.* [10]	Formulating the optimization of content router deployment as a convex optimization problem, and solving it by considering the trade-off between content router deployment cost and traffic transmission cost.
Energy-efficient content placement	Li *et al.* [11]	Establishing CCN energy consumption model, which relies largely on the average response hops of data dissemination.
	Choi *et al.* [12]	Considering different caching hardware technologies in CCN content router to reduce network energy consumption.
	Guan *et al.* [13]	Optimizing CCN content placement according to content popularity.
	Fang *et al.* [5]	Formulating energy consumption problem as a non-cooperative game, in which each content router makes local caching decisions considering both caching energy consumption and transport energy consumption.
	Llorca *et al.* [14]	Proposing an offline solution to maximize efficiency gains, and an distributed online solution to allow network nodes to make local caching decisions by estimating the current global energy benefit.
Energy-efficient request-to-cache routing	Kutscher *et al.* [9]	Proposing opportunistic caching and cache-aware routing techniques to gain the knowledge of a content's location.
	Sourlas *et al.* [15]	Proposing an intra-domain cache-aware routing scheme to minimize transportation cost based on the information item demands and the caching capabilities of the network.

Table 1. Energy-efficient caching techniques for information-centric networking.

Energy-Efficient Caching Techniques in Information-Centric Networking

Based on the analysis presented in the previous section, ICN in-network caching includes three important issues: cache placement (where is it best to put caches?), content placement (which content should go where?) and request-to-cache routing (how are cached content copies to be found?) [9]. For an ICN, the total consumed energy consists of two major parts: the transport energy and the caching energy. The transport energy includes the energy consumption in the core, edge, and access networks. The caching energy is consumed mainly by the contents cached in content routers, which obeys an energy-proportional model and depends on the caching hardware technology (e.g., SSD, DRAM, RLDRAM, SRAM, and TCAM). In this section, we give a summary of energy-efficient cache techniques in ICN from cache placement, content placement, and request-to-cache routing perspectives, which are presented in Table 1.

Energy-Efficient Cache Placement

ICN is a novel architecture that can be deployed to significantly reduce network capital cost at the lowest operating expense. The main hardware component of ICN is the content router with limited cache capacity. The position of ICN routers in a network has a direct influence on the overall performance. In this subsection, we address the practical issues regarding the possible deployment and evolution of ICN caches from an energy efficiency perspective.

Position of a Single Content Router — The position of a single router in ICN may have a significant impact on energy consumption. The further the content is cached from the user, the less duplication of content data is needed, and overall less storage energy is required. On the other hand, the closer the content is cached to the user, the faster the content can be served, and overall less transmission energy is required. For example, for different placement positions of the ICN router along the network described

in Fig. 4, Braun *et al.* [6] analyze the access frequency threshold from which caching content starts to be beneficial.

Core and Edge Deployment — As shown in Fig. 5, the feasibility of deploying ICN in edge and core networks can be investigated from both the cost and energy perspectives. By considering various incremental deployment scenarios (both core and edge), ICN can outperform conventional CDNs and P2P networks under the considered scenarios. For instance, with 20 percent deployment of CCN routers in the cores, CCN can effectively reduce the hop length, thereby reducing energy consumption more than 15 percent [2].

The optimization of content router deployment in large-scale information-centric core-edge separation Internet can be solved by considering the trade-off between content router deployment cost and traffic transmission cost. Chen *et al.* [10] assume that content routers are deployed randomly with a certain probability, and formulate the optimization problem as a convex optimization problem. Then they find that the optimal deployment probability is affected by the average number of hops for reaching the content provider, additional cost for building a content router, and traffic transmission cost. Moreover, for given content router deployment cost and traffic transmission cost, the optimal deployment probability for most content providers remains within a small range. In fact, it is more effective to deploy ICN nodes at the edge. For example, the scenario with 20 percent deployment in the edge routers performs almost as well as the scenario with 100 percent deployment in the core [2]. A caveat of this result is that the total number of CCN routers will be much greater when deploying them in the edge, because the number of edge routers tends to grow exponentially.

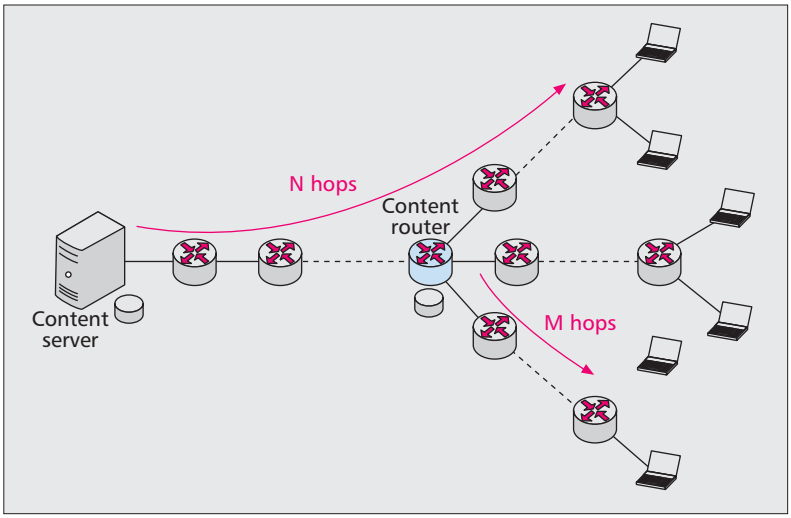


Figure 4. A network scenario of position of a single content router for ICN energy analysis.

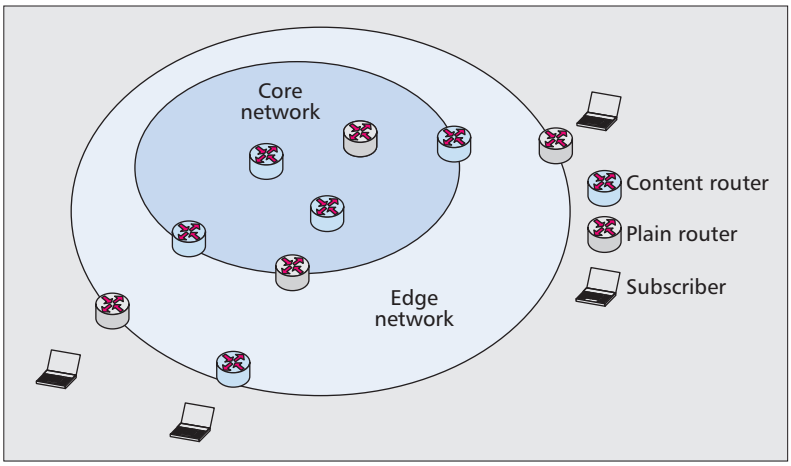


Figure 5. A network scenario of core and edge deployment for ICN energy analysis.

Energy-Efficient Content Placement — The placement of content replicas is another key issue for ICN in-network caching research. In an ICN network, content objects are dynamically created and requested, and can be cached as they travel toward end users, providing per-object-request granularity, responsiveness, and adaptation. The aim of the energy-efficient dynamic in-network caching problem is to find the evolution of the network configuration, in terms of the content objects being cached and transported over each network element at any given time, that meets user requests, satisfies network resource capacities, and minimizes overall energy use [14].

To maximize the energy saving, the energy optimization for ICN content delivery first needs to be designed. For example, an energy consumption model for CCN content delivery is studied in [11], where the energy consumption relies largely on the average response hops of data dissemination. Moreover, the authors of [12, 13] investigate the minimum energy consumption CCN can achieve with optimal content locations by considering different caching hardware technologies, number of downloads per hour, and content popularity. Although the optimization problem can be solved directly by solving in a centralized algorithm, all network information (e.g., the contents each node caches) in the network should be sent to a particular node to calculate the corresponding solutions, which should be distributed to the corresponding nodes in the network. Therefore, the centralized algorithm incurs huge communication overhead and lacks resilience to network changes.

To tackle the problem, the centralized energy consumption model can be transformed into a distributed one by adopting some proper tools. For instance, based on non-cooperative game, Fang *et al.* [5] propose an energy-efficient distributed in-network caching scheme to deal with the centralized energy consumption model for CCN, in which each content router only needs locally available information to make caching decisions considering both caching energy consumption and transport energy consumption. In addition, dual decomposition (DD) and the alternating direction method of multipliers (ADMM) methods can be considered. These transformed distributed energy consumption optimization algorithms can allow a global opti-

Object popularity is one of the major properties that affect cache efficiency. Research on current traffic patterns could shed additional light on the popularity characteristics of information today and thus benefit the design of energy-efficient caching schemes in ICN.

mization energy problem to be horizontally decomposed into parallel subproblems among the nodes.

To set the optimal solution as a benchmark, some distributed real-time caching policies based on network information (e.g., content popularity, user requests, equipment energy efficiency, and network topology) can also be designed to make a trade-off between the caching energy and the transport energy. For example, an efficient fully distributed online solution [14] is proposed to allow network nodes to make local caching decisions based on their current estimate of the global energy benefit.

Energy-Efficient Request-to-Cache Routing

In order to reduce energy consumption in ICN, requests have to be forwarded to the nodes that temporarily host (cache) the corresponding contents to take advantage of cached contents. This relates to energy-efficient request-to-cache routing, and the main challenge is that requests should ideally know the position of the cached content and follow the path to it. However, it is impractical to broadcast the instructions as to which content is cached where throughout the network. Therefore, the knowledge of a content's location at the time of the request might either not exist or be inaccurate (i.e., contents might have been removed by the time a request is redirected to a specific node).

To gain knowledge of a content's location, coordination between the data and control planes to update information of cached contents has been considered, but in this case scalability issues arise. There are two options to resolve the problem: opportunistic caching and cache-aware routing techniques [9]. In opportunistic caching, requests are forwarded to a server, and if the content is found on the path, the content is fetched from this node (instead of the original server). Cache-aware routing techniques can either involve both the control and data planes or only one of them. Furthermore, cache-aware routing can be done on a domain-wide scale or can involve more than an individual AS. For example, an intra-domain cache-aware routing scheme is proposed to compute the paths with the minimum transportation cost based on information item demands and the caching capabilities of the network [15]. In the latter case, business relationships between ASs might need to be exploited in order to build a scalable model.

Challenges and Future Research Directions

Although the caching technologies presented in the previous section can improve the energy efficiency of ICN, it is still a new research area full of challenges. There are a lot of issues that need to be addressed. In this section, we present some important yet challenging problems, and outline possible future research directions.

Chunk-Level Object Popularity — Object popularity is one of the major properties that affect cache efficiency. Research on current traffic patterns could shed additional light on the popularity characteristics of information today and thus benefit the design of energy-efficient caching schemes in ICN. In ICN, chunk-level object popularity rather than file-level object popularity should be considered. This line of study can be carried out in two directions. From the analytical point of view, the chunk-level object popularity model can be established from prior knowledge. These include established knowledge about file-level object popularity and distribution of object size, and reasonable assumptions on users' access behavior for chunks. From the experimental point of view, since currently there are no large-scale operational ICN network infrastructure and applications, it is difficult to measure the chunk-level object popularity directly. However, P2P systems, such as PPLive, can provide an opportunity to collect statistics about block-level object popularity. Certainly, the size of a block in P2P systems is different from the size of a chunk in ICN. However, their requesting behaviors are similar, so the results are analogous, and can be used at least as a reference for chunk-level object popularity in ICN.

Mix Traffic — Another issue is that when caching takes place in ICN, several types of traffic will compete for the same caching space. Therefore, cache space management becomes crucial for the network. Recent works, albeit based on simplified traffic models, have indicated that intelligent schemes can substantially improve energy efficiency in ICN. Although intelligent cache decision policies can reduce cache redundancy and increase the diversity of cached contents, making full use of content diversity needs complementary cache location mechanisms. How to devise and bundle low-complexity implicit cache decision policies and the corresponding intelligent cache location schemes in the face of the highly dynamic in-network cache environment remains an active research direction.

Wireless Information-Centric Networks — Mechanisms for energy-efficient caching have been studied mostly in the context of wired networks. With recent advances of wireless mobile communication technologies and devices, more and more end users access the Internet via mobile devices such as smart phones and tablets. This can create significant challenges in mobile environments, particularly mobile ad hoc networks (MANETs) and delay-tolerant networks (DTNs) due to the potential cost of managing cached replicas. Mobile node interests in content should be utilized to provide better network performance in terms of throughput, end-to-end delay, and energy consumption in both wireless and wired networks. Therefore, it is necessary to study the performance of existing energy-efficient caching techniques in ICN under wireless scenarios. Moreover, since currently there is no large-scale operational ICN network infrastructure and applications, large-scale tests on a real network (e.g., the PlanetLab environment) should be made to evaluate the actual effectiveness of these schemes under both wired and wireless scenarios.

Network Deployment — Another critical issue is that a sufficient number of content routers must be deployed throughout the network to reap the benefits of content caching. ICN can lower the energy consumption from the deployment and evolution perspective, but today's technology is not yet ready to support an Internet-scale CCN deployment. Nevertheless, by reducing the scope of a ICN deployment (i.e., from Internet scale to CDN or ISP scale), today's routers could easily be extended to become content routers. In this way, ICN can achieve energy efficiency while obviating the need to deploy preplanned and application-specific mechanisms, such as CDNs and P2P networks, which require sophisticated network services for mapping named content to hosts.

Although we consider equal deployment probability of content routers identified in the previous section, content routers can be deployed with various probabilities in different ASs. In addition, network topology, traffic flow, and the location of the content routers will also affect the traffic load of the information-centric core edge separation Internet. Therefore, it is interesting to investigate the optimization of content routers' deployment considering different deployment probabilities for core and edge network content routers. Moreover, investigating the deployment of content routers with consideration of the traffic flow and Internet topology is another research direction.

Conclusions

Information-centric networking is a novel networking architecture and promotes content to a first-class citizen in the network. In this article, we present the history and main components of ICN, and then introduce the architecture and workflow of the typical content-centric-networking-based ICN approach. Next, we present the issue of ICN in-network caching, and review some existing energy-efficient caching techniques. Finally, we outline some challenges and future research directions about energy-efficient caching policies for green ICNs.

References

[1] A. Bianzino *et al.*, "A Survey of Green Networking Research," *Commun. Surveys and Tutorials*, vol. 14, no. 1, 2012, pp. 3–20.

[2] U. Lee, I. Rimac, and V. Hilt, "Greening the Internet with Content-Centric Networking," *Proc. 1st ACM Int'l Conf. Energy-Efficient Computing and Networking*, Passau, Germany, Apr. 2010.

[3] U. Lee *et al.*, "Toward Energy-Efficient Content Dissemination," *IEEE Network*, vol. 25, no. 2, Mar. 2011, pp. 14–19.

[4] B. Mathieu *et al.*, "Information-Centric Networking: A Natural Design for Social Network Applications," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 44–51, July 2012.

[5] C. Fang *et al.*, "Energy-Efficient Distributed In-Network Caching for Content-Centric Networks," *Proc. IEEE INFOCOM '14 Wksps.*, Toronto, Canada, May 2014.

[6] T. Braun and T. Trinh, "Energy Efficiency Issues in Information-Centric Networking," *Proc. Euro. Conf. Energy Efficiency in Large Scale Distrib. Sys.*, Vienna, Austria, Apr. 2013.

[7] B. Ahlgren *et al.*, "A Survey of Information-Centric Networking," *IEEE Commun. Mag.*, vol. 50, no. 7, July 2012, pp. 26–36.

[8] G. Xylomenos *et al.*, "A Survey of Information-Centric Networking Research," *Commun. Surveys Tuts.*, no. 99, July 2013, pp. 1–26.

[9] D. Kutscher *et al*, "ICN Research Challenges," IRTF, Internet Draft, draft-kutscher-icnrg-challenges-00, Feb. 2013.

[10] J. Chen *et al.*, "Optimizing Content Routers Deployment in Large-Scale Information Centric Core-Edge Separation Internet," *Int'l. J. Commun. Sys.*, vol. 27, no. 5, May. 2012, pp. 794–810.

[11] J. Li, B. Liu, and H. Wu, "Energy-Efficient In-Network Caching for Content-Centric Networking," *IEEE Commun. Letters*, vol. 17, no. 4, Apr. 2013, pp. 797–800.

[12] N. Cho *et al.*, "In-Network Caching Effect on Optimal Energy Consumption in Content-Centric Networking," *Proc. IEEE ICC '12*, Ottawa, Canada, June 2012.

[13] K. Guan *et al.*, "On the Energy Efficiency of Content Delivery Architectures," *Proc. IEEE ICC'11 Wksps.*, Kyoto, Japan, June 2011.

[14] J. Llorca *et al.*, "Dynamic In-Network Caching for Energy Efficient Content Delivery," *Proc. IEEE INFOCOM '13*, Turin, Italy, Apr. 2013.

[15] V. Sourlas, *Replication Management and Cache Aware Routing in Information-Centric Networking*, Ph.D. thesis, Dept. ECE,, Univ. Thessaly, Greece, July 2013.

Since currently there is no large-scale operational ICN network infrastructure and applications, large scale tests on a real network (e.g. the PlanetLab environment) should be made to evaluate the actual effectiveness of these schemes under both wired and wireless scenarios.

Biographies

Chao Fang received his B.S degree in information engineering from Wuhan University of Technology, China, in 2009. He is currently working toward his Ph.D. degree at the State Key Laboratory of Networking and Switching Technology of Beijing University of Posts and Telecommunications (BUPT). Since August 2013 he has been visiting Carleton University, Ottawa, Ontario, Canada, as a visiting scholar. His current research interests include ICN, caching policies, energy-efficient resource management, and mobility support in ICN.

F. Richard Yu is an associate professor at Carleton University. He received the IEEE Outstanding Leadership Award in 2013, Carleton Research Achievement Award in 2012, Ontario Early Researcher Award (formerly Premier's Research Excellence Award) in 2011, Excellent Contribution Award at IEEE/IFIP TrustCom 2010, Leadership Opportunity Fund Award from Canada Foundation of Innovation in 2009, and Best Paper Awards at IEEE ICC '14, IEEE GLOBECOM '12, IEEE/IFIP TrustCom '09, and the International Conference on Networking '05. His research interests include cross-layer design, security, green IT, and QoS provisioning in wireless networks. He serves on the Editorial Boards of several journals, including *IEEE Transactions on Vehicular Technology* and *IEEE Communications Surveys and Tutorials*. He has served on the Technical Program Committees (TPCs) of numerous conferences, as TPC Co-Chair of IEEE GLOBECOM '14, IEEE INFOCOM-MCC '14, GLOBECOM '13, GreenCom '13, CCNC '13, INFOCOM-CCSES '12, ICC-GCN '12, VTC-Spring '12, GLOBECOM '11, INFOCOM-GCN '11, INFOCOM-CWCN '10, IEEE IWCMC '09, VTC-Fall '08, and WiN-ITS '07; and as Publication Chair of ICST QShine '10 and Co-Chair of ICUMT-CWCN '09.

Tao Huang received his B.S degree in communication engineering from Nankai University, Tianjin, China, in 2002, and M.S. and Ph.D. degrees in communication and information system from BUPT in 2004 and 2007, respectively. He is currently an associate professor at BUPT. His current research interests include network architecture, routing and forwarding, and network virtualization.

Jiang Liu received his B.S degree in electronics engineering from Beijing Institute of Technology in 2005, his M.S. degree in communication and information systems from Zhengzhou University, China, in 2009, and his Ph.D. degree in communication and information systems from BUPT in 2012. He is currently a lecturer at BUPT. His current research interests include network virtualization and future network architecture.

Yunjie Liu received his B.S degree in technical physics from Peking University, Beijing, China, in 1968. He is currently an academician of the China Academy of Engineering, chief of the science and technology committee of China Unicom, and dean of the School of Information and Communication Engineering, BUPT. His current research interests include next generation networks, network architecture, and management.

Approaches to Energy Intensity of the Internet

Dan Schien and Chris Preist

ABSTRACT

With more and more activities taking place online, concern over the environmental impact of digital services has drawn attention to the energy intensity of the network. Estimating the network energy intensity has been the subject of research for some time but results have differed widely, thus weakening the robustness of any conclusions drawn from assessments. A review of past studies shows two separate communities at work, applying different methods and assumptions. In this article we consider the approaches of top-down and bottom-up modeling. Top-down models have in the past usually given higher estimates of energy intensity than bottom-up models. We find that among the main reasons for the difference are varying system boundaries, and assumptions on the number and energy efficiency of routers and optical transmission equipment. Through application of consistent system boundaries around the metro and core networks and excluding access networks and customer equipment, we reduce the difference between the energy intensity estimates of the alternative approaches. Additionally, we review the varying assumptions in existing bottom-up models and combine them in a meta-model. Through Monte Carlo simulation over the distributions behind the varying assumptions we provide a more robust estimate of approximate energy efficiency for networks of 0.02 kWh/Gbyte that can be used in the environmental impact assessment of digital services.

INTRODUCTION

The continuing growth of digital services such as streaming videos, browsing websites or generally exchanging data over the Internet has drawn some attention to their environmental impact, which is either indirect, referring to the potentially beneficial impact of changes that digital services induce in the wider society and economy, or direct, resulting from manufacturing and energy consumption of devices. An understanding of the trade-offs between potential benefits and negative direct impacts enables consumers, businesses, and policy makers to take environmental impact into account.

Sustainability practitioners working for businesses providing digital services (e.g., online news or video) are experts in taking an end-to-end perspective and modeling all environmental impacts during the life cycle of a product, but lack the resources and expertise to create detailed models of each subsystem under consideration, such as the network. Instead, they require guidelines and off-the shelf models.

Despite some progress, efforts such as the information and communications technology (ICT) sector guidance service chapter to the Greenhouse Gas Protocol or the International Telecommunication Union (ITU) L 1410, "Methodology for the Assessment of the Environmental Impact of Information and Communication Technology Goods, Networks and Services" currently lack such models. Hence, practitioners adopt results from past studies without detailed analysis of underlying assumptions. In the case of energy usage by the Internet, this can be particularly problematic because the great variation of figures used means that the selection of one rather than another can dramatically affect the conclusions of an assessment.

In this article, we present a meta-analysis of past studies of energy consumption in the network. While this text provides an estimate of energy intensity of only edge, metro, and core networks, a complete assessment of a digital service needs to take all network parts into account, including the customer premises equipment (CPE), wired access networks, wireless access networks, and metro and long-haul networks, as shown in Fig. 1.

TWO COMMUNITIES: INDUSTRIAL ECOLOGY AND NETWORK RESEARCH

Life cycle assessments of digital services usually use a measure of the energy intensity of their network usage to determine their allocation of energy. This is normally stated in Joules per bit or kilowatt hours per gigabyte. It is calculated as a share of the network energy consumption relative to the data volume transported or, equivalently, of the power consumed per bandwidth sustained. Given the energy intensity, the energy footprint of a service is estimated as the product

The authors are with the University of Bristol.

0163-6804/14/$25.00 © 2014 IEEE

of energy intensity and the data volume of the service; per single unit of service (e.g., one minute of video stream) or for the entire audience (i.e., all videos streamed per year).

Energy intensity has been estimated using two different approaches: top-down and bottom-up. Each approach relates energy consumption to data traffic, but differs in the kind of input data it applies and enables its use in different applications. The distinction between top-down and bottom-up modeling approaches can be found in several domains where an overall property being calculated is also present on the level of model components.

A top-down model, for example [1], estimates the total energy use of an entire subsystem, such as "all data centers" or "the Internet," measures or estimates the total quantity of a given service type provided (e.g., data transmitted), and divides the former by the latter to give the energy consumption per unit of service. Hence, regarding energy consumption, it treats a given subsystem as a black box. Top-down models can evaluate change on the level of aggregate variables: total network traffic, average energy consumption per device class. Since top-down models are parameterized with market data, they are accessible to non-experts in network technology and are open to external validation.

For energy footprinting of digital services, the most influential top-down models (specifically, [1, 2]) were developed by researchers from the inter-disciplinary industrial ecology community — although this categorization is loose as academic communities are not clearly separated, and individual researchers publish in a variety of venues and collaborate across boundaries.

One of the defining goals of studies from this community is to quantify energy and material flows in industrial systems in order to increase the sustainability of industrial systems by understanding their relationships on technological, social, economic, and environmental levels. A particular focus is taking a whole systems perspective: investigating the dynamic relationship over these levels in order to prevent shifting the burden from one part of the system to another. For example, the shift toward distributed services provided by central servers through low-power clients might result in greater energy consumption by the network.

Top-down models are conducive to this whole systems perspective. Energy consumption in top-down models is usually estimated from market sales per device class and corresponding average power consumption to give a total over all considered device classes. By comparing this total with other macro-scale energy statistics, it is easy to sanity check them. And by treating the modeled system partly as a black box, they also do not require detailed knowledge on the network architecture. On the other hand, they cannot be used to evaluate changes to part of the network but only on trends of changing total network traffic or total energy consumption.

A bottom-up model (e.g., [3]), in contrast, calculates the overall energy intensity from the sum of the energy intensity of the subsystem components — usually the physical devices in the network. These models have also been referred to as transactional models as they allocate energy consumption to the transaction of data from end to end. As they represent energy intensity on the level of the system components, they are more flexible than top-down models to evaluate change: they can be used to evaluate the impact of modifications to the system architecture and its components.

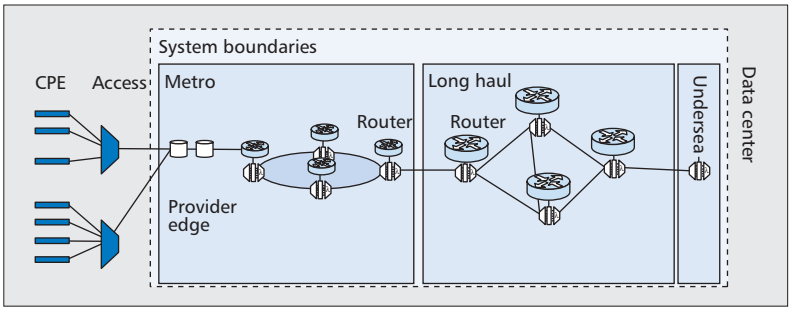


Figure 1. End-to-end model of the network. Servers in data centers are providing digital services to user devices in the home via the network, including CPE connecting to user devices in the home, access network, metro network, long- haul/core network, and undersea cable transport.

Such models thus require detailed knowledge of the operation and design of networks. Although this knowledge is already held by network operators, and thus, in principle, it is possible to represent each individual device in a bottom-up model, in practice network operators do not publicly disclose this information for business reasons. Instead, bottom-up models have been built based on implicit assumptions around the typical architecture of networks and are thus more difficult to validate.

It is no surprise that bottom-up models originate from the network research community, which investigates the design of networks and considers a number of metrics, including energy consumption. Network researchers have quantified energy consumption to estimate environmental impact from carbon emissions and, more frequently, to address network operator costs. If environmental impact was being assessed, the interpretation of results typically focused on directing research in network design.

Bottom-up models facilitate the evaluation of alternative design choices if they represent the network components that are to be altered. The scope of an investigation thus affects the level of detail at which the network is modeled. For example, if the goal of a study is to evaluate savings from optical switching, these devices must be explicitly modeled; if not, fiber optic components might be modeled in less detail with average values for energy consumption and capacity. At the same time the level of detail of modeling is naturally constrained by the simultaneously increasing complexity of the model, which is particularly relevant for end-to-end models.

Both modeling approaches introduce significant sources of uncertainty. The accuracy of a top-down model depends on the assumed total energy consumption and data volume. The accuracy of bottom-up models depends on how closely the assumed network architecture mirrors real network deployments. The energy elasticity of network devices, that is, the ratio between

Both modeling approaches introduce significant sources of uncertainty. The accuracy of a top-down model depends on the assumed total energy consumption and data volume. Accuracy of bottom-up models depends on how closely the assumed network architecture mirrors real network deployments.

marginal change of the utilization of a device and the resulting marginal change in energy consumption, cannot be taken into account by top-down models. But the current generation of network devices has very low energy elasticity, thus not limiting the potential accuracy of top-down models, as [4] found.

However, both approaches arrive at significantly different estimates for the energy intensity. In [5] a review of top-down and bottom-up models by Coroama and Hilty finds that top-down studies consistently arrive at higher estimates for energy intensity than bottom-up studies, differing by four orders of magnitude. Reasons for this discrepancy are given as being varying years of reference and varying system boundaries, sometimes including user devices, data centers, and CPE. However, they make no attempt to study the discrepancy by normalizing the system boundaries and thus leave open the question of whether the differences can be resolved. Table 1 and Fig. 1 list estimates of energy intensity of the past studies along a seemingly inverse exponential curve between 136 kWh/Gbyte and 0.006 kWh/Gbyte. We note that none of the three top-down studies with the highest energy intensity value includes end-user devices or optical fibers. Thus, even among studies with a similar year of reference and not including user devices a variation between one and two orders of magnitude remains. Specifically, the top-down model [2] and the bottom-up model [3], both with data for 2008 and both not including end-user devices or network CPE, arrive at estimates of 7 kWh/Gbyte and 0.006 kWh/Gbyte. Hence, the explanation for this discrepancy given by Coroama and Hilty is only partial, and further analysis is necessary to provide network energy intensity values for life cycle assessment practitioners.

In order to understand why bottom-up models arrive at significantly lower estimates than top-down variants, and to reduce the overall uncertainty, we review system boundaries of past models, and reconstruct the models within system boundaries around the edge and core network. We find that even with normalized system boundaries, the bottom-up models arrive at varying results due to different assumptions on bottom-up model parameters regarding the number and energy intensity of routers and fiber optical equipment. Without further qualification, it must be assumed that these models represent the real variability of existing network deployments. Based on this assumption, we then construct a bottom-up meta-model and parameterize it with distributions to represent the varying assumptions in existing models. By means of a Monte Carlo simulation we then generate a distribution of the overall energy intensity from which we suggest a new authoritative value for use by sustainability practitioners in assessments.

This article thus makes the following contributions:

- Provide a comparison of energy intensity estimates of top-down and bottom-up models within appropriate system boundaries for edge and core networks.
- Review the most robust bottom-up models of energy intensity and a normalization within common boundaries.

	2006 electricity (TWh/year)	
Equipment type	**Original inventory [1]**	**Reworked inventory**
Servers	24.5	
Data storage	4.4	
WAN switches	0.3	0.3
Routers	2.4	2.4
LAN switches	7.2	
Hub	3.5	
Transmission networks		1.2
Sum	42.3	3.9

Table 1. Inventory of annual energy consumption of system components in the top-down model [1] in the original system boundaries and our reworking.

- We present a distribution of the energy intensity with a single average value together with a confidence interval based on a principled approach.

Top-Down Models

The most influential top-down model comes from the industrial ecology community estimating the energy intensity of the U.S. Internet for 2006 (Taylor and Koomey [1]). In this study, the annual direct energy demand of the Internet is estimated as 19.3 TWh based on sales data of device types for 2000 in [6] and then extrapolated to a value of 42.3 TWh for 2006. A power usage effectiveness (PUE) value of 2 is applied additionally. The total energy consumption is then divided by an upper and lower bound estimate of annual network traffic of 5.4 to 9.6 Exabytes to give a resulting energy intensity of 9–16 kWh/Gbyte. The authors state that some assumptions were conservative, and thus the results constitute an overestimate. A later study [2] then applied an annual rate of reduction of 30 percent to the average between high and low estimates of energy intensity estimates of 2006 to account for increasing efficiency of devices and arrived at a value of 7 kWh/Gbyte for 2008. Extrapolated to 2014, this would result in a mean energy intensity of 0.84 kWh/Gbyte, a value that is considerably higher than most of the bottom-up estimates listed in the review [5].

To allow more accurate comparison of this model with bottom-up models of the core network, we rework this study to model the core network alone. We use the same data set [6] and methodology but change the system boundaries. Referring back to the end-to-end model of the network in Fig. 1, it is necessary to include fiber optic equipment, and edge and core routers and switches, but to exclude servers and data storage

Model variable				Mean (2014)	Unit
Single data points					
Overcapacity edge layer: 10,[2] PUE: 2,[1,2,3,4] Redundancy: 2[1,2,3,4]					
Triangular distributions	**Min**	**Mode**	**Max**	**Mean**	**Unit**
Energy intensity optic amplifiers	0.03 (0.065)[2]	0.21 (0.27)[3]	Proprietary (Proprietary)[4]	Proprietary	J/Gb
Energy intensity optical switch	0.04 (0.05)[3]	0.35 (0.46)[2]	1.42 (1.85)[2]	0.60	J/Gb
Number core hops	3[2]	6[4]	10[1]	6.33	—
Number metro hops	3[1]	4[2]	12[4]	6.33	—
Total distance	7500[1]	7500[2]	8217[4]	7739	km
Uniform distributions	**Lower**	**Upper**	**Mean**	**Unit**	
Energy intensity regenerator	7.66 (10)[3]	Proprietary (Proprietary)[4]		Proprietary	J/Gb
Distance sea cable	6000	12000[4]		9000	km
Energy intensity edge switch	2 (4.46)[1]	3.59 (8)[2]		2.80	J/Gb
Energy intensity OTN switch	1.57 (3.5)[2]	2.6 (3.4)[3]		2.09	J/Gb
Energy intensity per km undersea cable	0.021[8]	0.066[9]		300	J/Gb/km
Energy intensity transponder	2.11 (4.7)[2]	3.83 (5)[3]		2.97	J/Gb
Overcapacity core layer	2[1,2]	4[4]		3	—
Overcapacity metro layer	5[2]	10[6]		7.50	—
Span optical amplifier	80[2]	100[1]		90	km
Undersea traffic share	0.1[10]	0.5[10]		0.3	—

[1] [3], [2] [7], [3] [11], [4] [10], [5] Exclude regenerators from model,
[6] Half of the utilization in the SWITCH research network reported in [10],
[7] Based on the total power consumption and number of optical amplifiers and regenerators for the Internet2 core network from conversation with the authors of [10].
[8] Based on formula 15 in [3].
[9] Based on the average energy intensity per km in [10].
[10] Portion of undersea traffic varies with location of the user and service.

Table 2. Model parameters including parameter name, the type of distribution applied in the Monte Carlo simulation, and the distribution parameters. Parameters are grouped by applied distribution type: uniform, triangular, choice. For a uniform distribution, min and max denote the boundary values. A choice denotes discrete values. A point estimate refers to a single value. Device energy intensity values are listed with their extrapolated value for 2014 based on a 12.5 percent annual improvement rate with the original value in brackets. References to the original sources are listed at the bottom of the table and are referred to by superscript indices.

as well as campus network equipment such as office floor hubs and small switches. While this is mostly straightforward, the router category includes both high-end core routers and small-office-level models. Given that the router category is the largest position in the inventory, this results in a significant overestimate. Although the two remaining top-down models in [5] are not based on Roth's inventory, these cannot be used to triangulate the portion of core routers from all routers as one is equally focused on campus networks, and the other only provides an aggregate result for network device energy consumption.

In Table 1 we list the inventory categories from [6] as used in [1] and in our reworking to focus on the core network alone. The resulting estimate of total annual energy consumption in the updated inventory is 3.9 TWh/year compared

Any variation in the overall energy intensity results from different assumptions on the route length of metro and core networks as well as the energy intensity of router and fiber optic equipment, which varies between specific device types and models and with device age.

to 42.3 TWh/year in the original estimate. If the assumed 30 percent annual improvement rate by Weber and colleagues in [2] is applied to this estimate, the resulting energy intensity would be 0.55 kWh/Gbyte for 2009 (down from 7 kWh/Gbyte) and 0.07 kWh/Gbyte for 2014 (down from 0.84 kWh/Gbyte). However, this annual improvement rate was calculated relative to the observed growth of energy consumption by data center network equipment from 2000 to 2006, and might be too high for carrier network equipment. Kilper *et al.* [7] refer to Tamm *et al.* [8] for an estimation of annual improvements of telecom equipment of 10 percent. Although Tamm *et al.* do not provide the value of 10 percent explicitly, a reconstruction of their Fig. 7 results in a value of 12.5 percent. At this lower annual improvement, the resulting average energy intensity for the top-down model would be 0.39 kWh/Gbyte for 2014.

More recently, another top-down model for the Swedish core network by Teliasonera estimated its energy efficiency as 0.08 kWh/Gbyte for 2010 [9] which is lower but not entirely dissimilar to our reworked values. This study is supported by confidential data from Teliasonera and thus important for corroboration, but does not provide enough detail in order to compare and explain differences to other studies.

Bottom-Up Models

Bottom-up models of end-to-end network energy intensity combine a network architecture for the access, metro, and core network layers, with a specific parameterization of device energy intensity values.

Any variation in the overall energy intensity results from different assumptions on the route length of metro and core networks as well as the energy intensity of router and fiber optic equip-

Router model	Source	Energy intensity (J/Gb)	Year of reference	Energy intensity 2014 (J/Gb)
Metro routers				
Cisco 12816	[3]	25.75	2008	11.56
Cisco 7603	[10]	25.00	2009	12.82
Cisco 7606	[10]	16.04	2009	8.23
Cisco 7613	[3]	38.33	2008	17.20
Cisco 10008	[3]	137.50	2008	61.71
Hitachi GS4000 320E	[10]	12.50	2009	6.41
Hitachi GS4000 160E	[10]	10.00	2009	5.13
Cisco 6513	[3]	8.36	2008	3.75
Cisco 6513	[10]	40.00	2009	20.52
Cisco 6509	[10]	40.00	2009	20.52
Juniper MX960	[10]	16.20	2009	8.31
Mean Energy Intensity [J/Gb]		33.61		16.01
Core routers				
Juniper T1600	[10]	34.48	2009	17.69
Juniper T640	[10]	17.47	2009	8.96
Juniper T320	[10]	16.20	2009	8.31
Cisco CRS – 1	[3]	17.03	2008	7.64
Cisco CRS – 3	[11]	10.00	2012	7.66
Generic	[7]	12.60	2008	5.65
Mean energy intensity (J/Gb)		17.96		9.32

Table 3. Energy intensity of metro and core routers as provided in previous studies and extrapolated to 2014 based on an improvement rate of 12.5 percent per year.

ment, which varies between specific device types and models, and with device age. Additionally, overheads for building infrastructure, expressed as PUE, redundancy, and overcapacity, increase the overall network energy intensity.

One of the first end-to-end models, by Baliga and colleagues [3], estimated power draw per user of the optical Internet as a function of bandwidth in the access network for several access network technologies and has been referenced in assessments of digital services several times. Given that user bandwidth was estimated from statistical average values, the overall estimate can be converted equivalently to energy consumption per bit. More recent formulations of the model by the same authors have maintained parameterization and architecture largely unaltered.

They describe a reference end-to-end architecture for the network including CPE, access, edge, metro, and long-haul networks, undersea cables, as well as an additional IPTV network.

Baliga and colleagues only published the energy efficiency result including the access network, which depends on the access rate. Our reproduction of their model resulted in a value for the energy consumption of the core Internet — not including the access network but including undersea traffic — of 2.66 J/Mb, which equals 0.0059kWh/Gbyte. As the authors acknowledge in their text, the model provides an underestimate of the energy efficiency of the Internet.

Another notable end-to-end bottom-up model by Kilper *et al.* [7] evaluates how the power consumption of optical networks is likely to change through 2020 and take a mix of different types of services into account. They provide a layered network path model to estimate energy consumption for services using a specific network topology (e.g., peer-to-peer vs. video) by summing up the energy consumption of each layer traversed. However, unlike Baliga, they do not include undersea cables and associated terminals in their model.

In total, our reproduction of their model yields an energy consumption of 3.28 J/Mb (0.0073 kWh/Gbyte) for a path that includes one leg of edge, metro, and long haul networks, which despite the absence of a leg of undersea cable is higher but of comparable magnitude to Baliga's values. The year of reference for equipment efficiency values in both studies is 2008 [5, 7].

The distribution of energy intensity over the subsystems is substantially different in the two models. Although Kilper *et al.* agree with Baliga *et al.* that the core layer is more impactful than the edge, their model indicates that the fiber optic devices contribute to a much greater degree.

Although we consider these studies to be most robust end-to-end models of energy intensity, there are many other excellent models of energy consumption in networks. However, these usually model energy consumption on a network scale (as opposed to end-to-end) or evaluate relative changes without providing absolute values, and thus are not applicable to our needs. Reference [12] provides the most detailed model of the optical layer, thus providing a valuable source of individual parameters and corroborating assumptions in the meta model we present below.

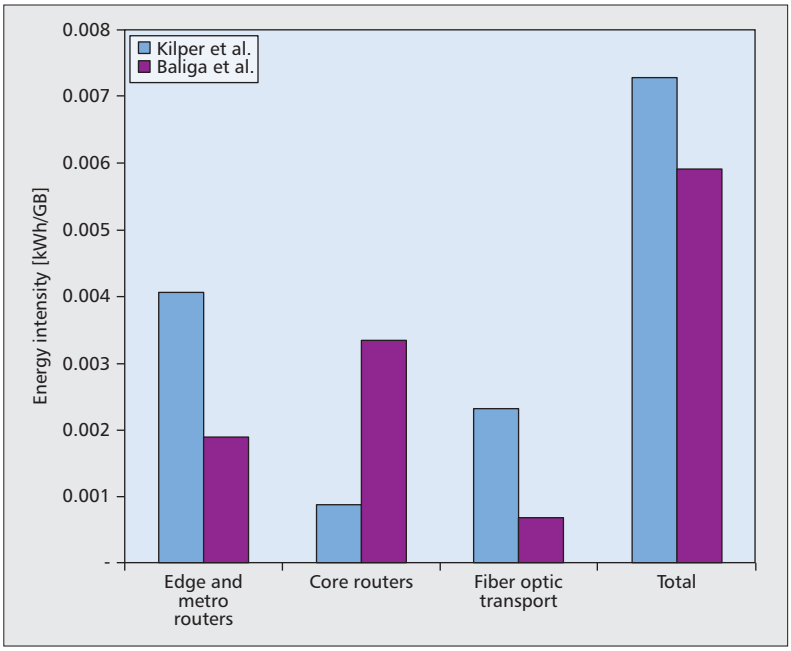


Figure 2. Composition of energy intensity by edge and metro routers, core routers, and fiber optic transport in Baliga *et al.* [3] and Kilper *et al.* [7]. The fiber optic transport includes overland and subsea cables in [3].

Finally, a study from the industrial ecology community by Coroama *et al.* [10] provides an estimate of the energy intensity of transporting the video signal of a virtual conference on a network path between Switzerland and Japan of 0.2 kWh/Gbyte, which is higher than our reworked top-down estimate. This study is unique in that network operators provided specific values for power consumption, and utilization and capacity of routers and fiber optic transmission equipment, which is relevant to validate other bottom-up models from the network research community. As the authors acknowledge, the study investigates a worst case scenario given the unusually long distance of the video channel spanning three continents from Europe across the United States to Japan, around 27,000 km of distance.

Meta-Model

In the highest estimates of energy intensity for edge and core networks by the bottom-up models in the previous section 0.2 kWh/Gbyte is 33 times higher than the lowest of 0.0059 kWh/Gbyte. Although the route between Davos and Nagoya analyzed by Coroama and Hilty is untypically long, this difference alone cannot explain the variance. Given that there is no clear underlying reason in the models to favor one over the other, the difference partly represents the actual variability found in real network deployments, and the actual average energy intensity of edge and core networks is to be found along the spectrum defined by the variability of underlying parameters. We now combine different data and structural assumptions

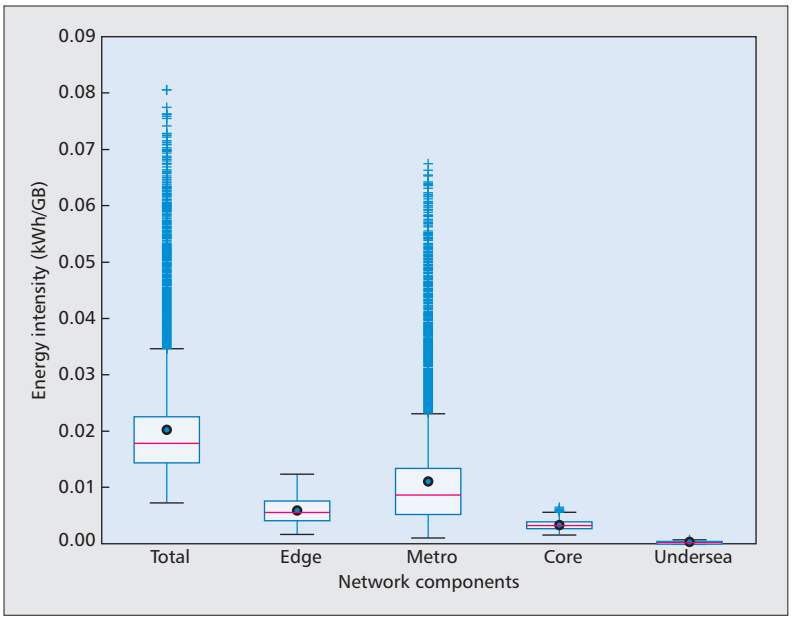


Figure 3. Box and whisker plot of energy intensity from the Monte Carlo simulation of the meta-model showing the total energy intensity for the edge and core networks, and the energy intensity for individual layers. The red line indicates the median values, the vertical edges of the boxes mark the first (lower edge) and third (upper edge) quartiles. The blue dots indicate the mean. The horizontal black lines indicate 1.5× the inter quartile range (IQR), the distance between the first and third quartiles. Outliers outside of the IQR are marked as crosses.

within the models discussed so far to allow us to calculate an average value for the energy intensity of the core Internet. From the studies discussed above, we adopt the most detailed and robust model for each layer under consideration. Similar to [7], we model each layer as composed of a number of nodes (IP + fiber optic devices), and the energy intensity per layer is the sum of the energy intensity of all nodes in a layer multiplied by factors for overcapacity, PUE, and redundancy. These intensity values are then added over all devices that constitute a layer. The overall energy intensity is the sum of the edge, metro, and core layers. We follow [11] in explicitly modeling the components of the optical layer (optical transport network switches, transponders, line amplifiers, regenerators), and we follow [3] in the modeling of the undersea transport. We apply an energy efficiency improvement rate of 12.5 percent per annum (taken from [8]) on deployed network devices to normalize all data to a reference year of 2014. The model is available in code and with typeset documentation online.[1]

Given the combined structure and parameterization of the model, we then perform a Monte Carlo simulation to give us a distribution of the overall energy estimate including a mean value to represent the average case of core networks in general.

The parameterization of the model is provided in Table 2. The distributions for routers are calculated by resampling from a Gaussian kernel density estimated distribution, using the data in Table 3. For PUE and redundancy, we apply the same single value of 2 that was assumed by all studies.

[1] http://nbviewer.ipython.org/gist/dschien/1859c0f525473211f66f.

The distribution of hops in metro and core networks is based on the assumptions in [3, 7, 10]. In [10] 6 hops are located in the core network, and 12 hops are in the Swiss and Japanese research networks SWITCH and NICT with 7 and 5 hops, respectively, which we use as the high estimate for the metro network.

Results from the bottom-up model are strongly influenced by assumptions of the network utilization or overcapacity. This refers to the difference between maximum capacity, which serves as the basis for the calculation of the devices' energy intensity, and the actual use of capacity. Reference [3] assumes no overcapacity for edge and metro, which is an idealization we ignore. In [10] utilization on routers and links in the core network combined is 26.3 percent, excluding the undersea cables and terminals, resulting in an overhead coefficient of 4. Reference [10] also provides utilization values for the SWITCH research network of 5 percent, which we exclude because it is likely to be lower than commercial networks.

Results

The resulting distribution from the Monte Carlo simulation is displayed as a box and whisker plot in Fig. 3 showing the total energy intensity as well as the contribution of the edge, metro, and core layers, and undersea segments.

The mean energy intensity for 2014 is 0.02 kWh/Gbyte with 25th and 75th percentiles of 0.0144 and 0.023, respectively, and a median of 0.18 kWh/Gbyte.

In Fig. 4 we compare the energy intensity values of the studies discussed so far with that of the reworked top-down model (0.39 kWh/Gbyte) and the bottom-up meta-model (0.02 kWh/Gbyte). Although the discrepancy is substantially reduced from the original estimates, they cannot be compared like for like due to the inclusion of campus-level routers in the top-down estimate, which highlights an important area for further research. Other reasons that will contribute to the discrepancy will be:

- The age of the underlying data behind the top-down estimate means that the margin of error of the projection forwards is high.
- Bottom-up models tend to be leaner, and will miss some deliberate redundancy or spare equipment.

The top-down model by Malmodin *et al.* [9] with 0.08 kWh/Gbyte arrives at a value that is only four times higher than the meta-model result and thus provides partial corroboration. Unfortunately, it does not provide a detailed account of the model inventory to investigate what specific assumptions differed or were identical.

Discussion and Conclusion

The estimate of resulting energy intensity presented in the previous section is based on normalizing boundaries and statistically combining assumptions from previous studies, which in turn were based on measurements and experience.

The overall estimate can be used in sustainability assessments to estimate the network ener-

gy consumption that can be attributed to specific digital services. For example, a content service provider such as the BBC could apply this energy intensity to estimate the network energy consumption to be attributed to the downloading one hour of HD video on the BBC iPlayer service. The associated file is approximately 1 Gbyte in size, and so would be attributed 20Wh of the energy consumed by the core and edge networks. More detailed analyses of this kind can be used to explore the impact on energy consumption of alternative deployment architectures for digital services [12].

A complete model of energy consumption involved in the delivery of digital services must also model access networks — the use of home, campus, and mobile networks to access a service. The energy consumption of these is significant. As we argue in [12], usage characteristics of such equipment means that energy intensity is not an appropriate metric, and other allocation approaches are needed. This is discussed further in [13], and models are proposed.

More broadly, energy intensity of the network constitutes an example of an assessment of environmental impact of an industrial system. For these to be reliable, more input from the engineering community is required. Models by the industrial ecology community tend to provide overestimates, in order to err on the safe side, while engineering models tend to provide underestimates, for example, by abstracting from legacy systems where not needed.

Further research is necessary to provide a transparent inventory for top-down models that specifically identifies service provider network routers from campus network routers.

Both the community of network researchers and that of industrial ecology can contribute in order that the energy intensity values, which are continuously used by practitioners, are reliable and accurate.

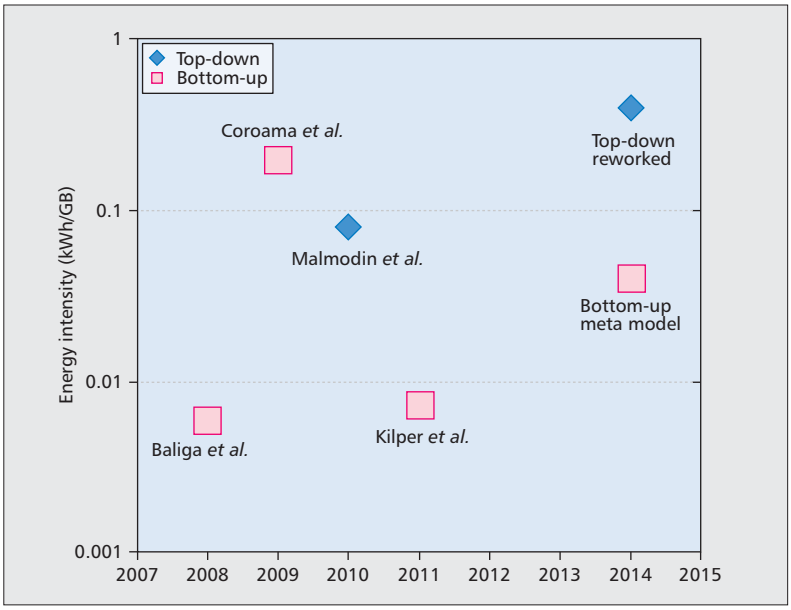


Figure 4. Energy intensity estimates from top-down and bottom-up models with similar system boundaries on a log scale.

Acknowledgment

This work was funded in part by UK Research Council Digital Economy grant SYMPACT (grant number EP/I000151/1) as well as partly by UK EPSRC Large-Scale Complex IT Systems Initiative (grant EP/F001096/1).

References

[1] C. Taylor and J. G. Koomey, "Estimating Energy Use and Greenhouse Gas Emissions of Internet Advertising," working paper for IMC2, 2008.
[2] C. L. Weber, J. G. Koomey, and H. S. Matthews, "The Energy and Climate Change Impacts of Different Music Delivery Methods," 2009.
[3] J. Baliga *et al.*, "Energy Consumption in Optical IP Networks," *J. Lightwave Tech.*, vol. 27, no. 13, July 2009, pp. 2391–2403.
[4] C. A. Chan *et al.*, "Methodologies for Assessing the Use-Phase Power Consumption and Greenhouse Gas Emissions of Telecommunications Network Services," *Environ. Sci. Tech.*, Dec. 2012.
[5] V. C. Coroama and L. M. Hilty, "Assessing Internet Energy Intensity: A Review of Methods and Results," *Environ. Impact Assess. Rev.*, vol. 45, Feb. 2014, pp. 63–68.
[6] K. W. Roth, F. Goldstein, and J. Kleinman, *Energy Consumption by Office and Telecommunications Equipment in Commercial Buildings Volume I: Energy Consumption Baseline*, 2002.
[7] D. C. Kilper *et al.*, "Power Trends in Communication Networks," *IEEE J. Quantum Electron.*, vol. 17, no. 2, 2011, pp. 275–84.
[8] O. Tamm and C. Hermsmeyer, "Eco-Sustainable System and Network Architectures for Future Transport Networks," *Bell Labs Tech.*, vol. 14, no. 4, 2010, pp. 311–27.
[9] J. Malmodin *et al.*, "Life Cycle Assessment of ICT," *J. Ind. Ecol.*, May 2014.
[10] V. C. Coroama *et al.*, "The Direct Energy Demand of Internet Data Flows," *J. Ind. Ecol.*, vol. 17, no. 5, July 2013, pp. 680–88.
[11] W. Van Heddeghem *et al.*, "Power Consumption Modeling in Optical Multilayer Networks," *Photonic Net. Commun.*, vol. 32, Jan. 2012.
[12] D. Schien *et al.*, "Modeling and Assessing Variability in Energy Consumption During the Use Stage of Online Multimedia Services," *J. Ind. Ecol.*, vol. 17, no. 6, Dec. 2013, pp. 800–13.
[13] V. C. Coroama *et al.*, "The Energy Intensity of the Internet: Home and Access Networks," *ICT Innovations for Sustainability*, vol. 310, L. M. Hilty and B. Aebischer, Eds., 2015, pp. 137–55.

Biographies

Daniel Schien (daniel.schien@bristol.ac.uk) received his Diplom (M.Sc. equivalent) in computer science from TU Berlin, Germany. He is currently working toward a Ph.D. degree in computer science at the University of Bristol, United Kingdom. His research interests include environmental assessments of ICT and smart city solutions.

Chris Preist (Chris.Preist@bristol.ac.uk) is a reader in sustainability and computing systems at the University of Bristol. He was principal investigator on the SYMPACT project, working with Guardian News and Media on how the digital transformation of the news and media sector will impact energy use, greenhouse gas emissions, and other sustainability factors. Prior to joining the University of Bristol, he was head of Sustainable IT Research at HP Labs, Bristol, where he led work on the strategic impact of climate change on business and technology development to exploit emerging opportunities.

ACCEPTED FROM OPEN CALL

Assessing and Safeguarding Network Resilience to Nodal Attacks

Pin-Yu Chen and Alfred O. Hero III

ABSTRACT

This article introduces new methods for evaluating and improving resilience of network connectivity to attacks on nodes of the network. Network connectivity is evaluated using a centrality measure that quantifies sensitivity of the size of the largest connected component to node removals. Based on this centrality measure, a new method for improving resilience is introduced, called edge rewiring. The topology of the power grid of western U.S. states is used to illustrate the proposed method. Using the proposed centrality measure, we show that the power grid topology is especially vulnerable to nodal attacks. In particular, using the proposed centrality measure, an attacker could reduce the largest component size by nearly a factor of two by only targeting 0.2 percent of the nodes. More importantly, we show that network resilience can be greatly improved via a few edge rewires without introducing additional edges in the network.

INTRODUCTION

The problem of establishing resilience of network connectivity to node removals has received much recent attention [1–5]. Resilience is closely related to reliability of networks when a subset of nodes are inactivated. It arises in applications including service disruption in communication systems caused by router failures, and blackout in power systems caused by power station shutdowns, among others. In these applications network functionality can be disrupted by targeted attacks, for example, denial of service (DoS) or jamming attacks, or by natural occurrences, such as weather-related link failures and power outages. In this article we introduce a new method for assessing the resilience of networks to node removals and preventive approaches to desensitize numerous connectivity attacks.

A resilient network has global connectivity and largest component size that are only minimally disrupted by limited attacks on nodes or edges. For example, a fully connected network allows communication between all pairs of nodes, and its largest component is the entire set of nodes in the network. One measure of network connectivity is given by the standard graph-theoretic k-connectivity definition: a graph is k-connected if any set of $k - 1$ node removals does not disconnect the graph. However, this definition does not account for the number of communication paths between nodes that are disrupted, which is more relevant to the functioning of the network. A more relevant measure of connectivity is proposed here: the minimum number of node removals necessary to reduce the size of the largest component by a fixed proportion (e.g., 10 or 50 percent) of its original size.

To illustrate, consider a large network where one of its nodes is connected to the rest of the network by a single edge (i.e., node degree one). Removing this edge (or the adjacent node) will reduce both the number of communication paths and the largest component size by one. However, if the network is composed of two cliques of equal size connected by a single edge, removal of this edge will reduce the number of paths and the largest component size by a factor of two.

A node centrality measure is a quantity that measures the level of importance of a node in a network. The utility of centrality measures is that they can break the combinatorial bottleneck of searching through all the possible permutations and combinations of nodes that might reduce the largest component size. An attack that removes nodes according to a measure of centrality, such as the one introduced in the next section, is referred to as a *centrality attack*. For example, the authors of [1–4] study the effectiveness of degree centrality attacks (i.e., removing the largest hub nodes) as a way to reduce the size of the largest component of the network. However, it has been shown in [5] that node degree is not the most effective centrality measure for minimizing largest component size. For different network topologies, investigating resilience of network connectivity to centrality attacks provides a unified metric for evaluating network vulnerabilities.

Quantitative network resilience measures can also be used to assess the effectiveness of preventive approaches for hardening a network against attacks. Two preventive approaches are discussed in this article. The first method is the *edge addition* method [6], where edges are added to the network to enhance network resilience. The second method is the proposed *edge rewiring* method, where new edges are introduced by swapping a subset of existing edges.

The authors are with the University of Michigan.

0163-6804/14/$25.00 © 2014 IEEE

One possible advantage of edge rewiring over edge addition is that edge rewiring requires no additional edges to enhance network resilience. The edge rewiring method might be preferable to the edge addition method in the following aspects:

- *Lower operational and maintenance costs*: For power grids, power dissipation and facility maintenance costs are proportional to the total number of edges in the network.
- *Easier link monitoring for network security*: In large-scale systems such as the Internet and cellular infrastructures, introducing additional edges inevitably raises the security risks to information exposure, and also incurs extra burden for system administration and monitoring.
- *Reduced provisioning budget*: In networking applications with stringent energy/bandwidth constraints, such as sensor networks and peer-to-peer (P2P) networks, introducing additional edges consumes more networking resources.

To illustrate resilience of network connectivity to different centrality attacks, and effectiveness of preventive approaches, we consider the power grid network for western U.S. states [8]. We show that different centrality measures differ significantly in their ability to assess resilience of this real-world network. If the proposed centrality measure is used by an attacker, the largest component size can be reduced to nearly half of its original size by removing only 0.2 percent of nodes in the network. Attacks using other types of centrality measures are less effective in reducing largest component size. In particular, even if as many as 1 percent of the nodes are removed, less than 6 percent reduction in largest component size is achieved by other types of centrality attacks. In addition, we show that the proposed edge rewiring method can greatly improve network resilience via only a few edge rewires while achieving the same performance as the edge addition method. A second illustrative example for a European Internet backbone network is discussed in the supplementary file.[1]

The rest of the article is organized as follows. The next section reviews several centrality measures summarized in Table 1. The section following that investigates the resilience of network connectivity to different centrality attacks on the power grid topology. The article then provides a discussion of the edge addition method and the proposed edge rewiring method as preventive approaches to centrality attacks. Next, we implement the two preventive approaches and evaluate their performance on the power grid topology. Finally, the last section concludes the article. We adopt the following notation conventions. Uppercase letters in boldface represent matrices, lowercase letters in boldface represent vectors, and uppercase letters in calligraphic face represent sets. $(\cdot)^T$ denote matrix and vector transpose.

Centrality Measures

A network is a connected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. The connectivity structure of G can be represented by the $|\mathcal{V}| \times |\mathcal{V}|$ adjacency matrix $\mathbf{A}$, where $|\mathcal{V}|$ is the number of nodes in G and $\mathbf{A}_{ij} = 1$ if nodes i and j are connected by an edge; otherwise, $\mathbf{A}_{ij} = 0$. Let $\mathcal{N}_i$ denote the set of nodes connecting to node i (i.e., the set of neighbors of node i), and let $|\mathcal{N}_i|$ denote the set size. The degree of node i is the number of edges connected to it, that is, $d_i = \Sigma_{j=1}^{|\mathcal{V}|} \mathbf{A}_{ij} = |\mathcal{N}_i|$. The degree matrix $\mathbf{D}$ is defined as $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_{|\mathcal{V}|})$, where $\mathbf{D}$ is a diagonal matrix with degree information on its main diagonal with the rest of the entries being 0. The graph and Laplacian matrix $\mathbf{L}$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, and therefore it encodes degree information and connectivity structure of a graph. $\mathbf{L}$ is a positive semidefinite matrix, all its eigenvalues are nonnegative, and $\text{trace}(\mathbf{L}) = 2|\mathcal{E}|$, where $\text{trace}(\mathbf{L})$ is the sum of eigenvalues of $\mathbf{L}$ and $|\mathcal{E}|$ is the number of edges in G. Moreover, the smallest eigenvalue of $\mathbf{L}$ is always 0, and the eigenvector of the smallest eigenvalue is a constant vector. The second smallest eigenvalue of $\mathbf{L}$, denoted by $\mu(\mathbf{L})$, is also known as the algebraic connectivity [9]. It has been proven in [9] that $\mu(\mathbf{L})$ is a lower bound on node and edge connectivity for any non-complete graph. That is, algebraic connectivity ≤ node connectivity ≤ edge connectivity.

	Global measure	Local measure	Mathematical expression
Betweenness	✓		$\text{betweenness}(i) = \Sigma_{k\neq i}\Sigma_{j\neq i,j>k}(\sigma_{kj}(i))/(\sigma_{kj})$
Closeness	✓		$\text{closeness}(i) = 1/\Sigma_{j\in\mathcal{V},j\neq i}\rho(i,j)$
Eigenvector centrality (eigen centrality)	✓		$\text{eigen}(i) = \lambda_{\max}^{-1}\Sigma_{j\in\mathcal{V}}\mathbf{A}_{ij}\xi_j$
Degree		✓	$d_i = \Sigma_{j=1}^{\lvert\mathcal{V}\rvert}\ \mathbf{A}_{ij}$
Ego centrality		✓	$\text{ego}(i) = \Sigma_k\Sigma_{j>k}1/[\mathbf{A}^2(i)\circ(\mathbf{I}-\mathbf{A}(i))]_{kj}$
Local Fiedler vector centrality (LFVC)	✓[2]		$\text{LFVC}(i) = \Sigma_{j\in\mathcal{N}_i}(y_i - y_j)^2$

Table 1. Summary of centrality measures and their properties.

The centrality of a node is a measure of the node's importance to the network. Centrality measures can be classified into two categories, *global* and *local* measures. Global centrality measures require complete topological information for their computation, whereas local centrality measures only require partial topological information from neighboring nodes. For instance, acquiring shortest path information between every node pair is a global method required for the betweenness centrality measure, and acquiring degree information of every node is a local method. Some commonly used centrality measures are:

Betweenness [10]: Betweenness is the fraction of shortest paths passing through a node relative

[1] See supplementary file available at http://sites.google.com/site/pinyuchenpage/publications.

[2] Although LFVC is a global centrality measure, it is locally computable via distributed power iteration method [7].

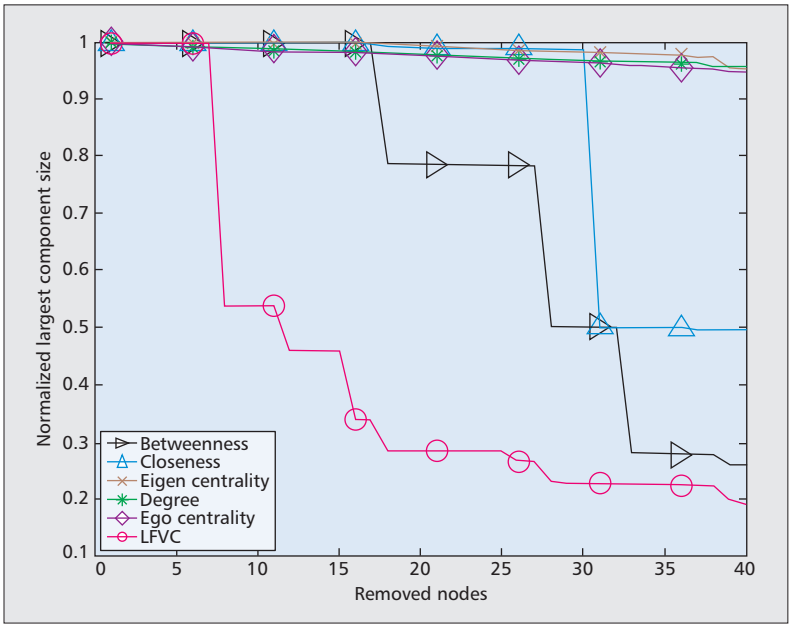


Figure 1. Resilience of network connectivity to different centrality attacks on the power grid topology of western U.S. states [8]. This network contains 4941 nodes and 6594 edges, where nodes represent power stations and edges represent power lines. By removing roughly 0.2 percent of the nodes in the network based on an LFVC attack, the largest component size is reduced to nearly half of its original size.

to the total number of shortest paths in the network. Specifically, it is a global measure defined as

$$\text{betweenness}(i) = \sum_{k \neq i} \sum_{j \neq i, j > k} \frac{\sigma_{kj}(i)}{\sigma_{kj}},$$

where σ_{kj} is the total number of shortest paths from k to j, and $\sigma_{kj}(i)$ is the number of such shortest paths passing through i.

Closeness [11]: Closeness is a global measure of shortest path distance of a node to all other nodes. A node is said to have higher closeness if the sum of its shortest path distance to all other nodes is smaller. Let $\rho(i,j)$ denote the shortest path distance between nodes i and j in a connected graph. $\text{closeness}(i) = 1/\Sigma_{j \in \mathcal{V}, j \neq i} \rho(i,j)$.

Eigenvector centrality (eigen centrality): Eigenvector centrality depends on the ith entry of the eigenvector associated with the largest eigenvalue of the adjacency matrix $\mathbf{A}$. It is defined as $\text{eigen}(i) = \lambda_{\max}^{-1} \Sigma_{j \in \mathcal{V}} \mathbf{A}_{ij} \xi_j$, where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{A}$ and ξ is the eigenvector associated with $\lambda_{\max}$. It is a global measure since the eigenvalue decomposition of $\mathbf{A}$ requires complete topological information of the entire network.

Degree (d_i): Degree is the simplest local centrality measure and is simply the number of neighboring nodes.

Ego centrality [12]: Consider the $(d_i + 1)$-by-$(d_i + 1)$ local adjacency matrix of node i, denoted by $\mathbf{A}(i)$, and let $\mathbf{I}$ be an identity matrix. Ego centrality can be viewed as a local version of betweenness that computes the shortest paths between its neighboring nodes. Since $[\mathbf{A}^2(i)]_{kj}$ is the number of two-hop walks between k and j, and $[\mathbf{A}^2(i) \circ (\mathbf{I} - \mathbf{A}(i))]_{kj}$ is the total number of two-hop shortest paths between k and j for all $k \neq j$, where $\circ$ denotes the entrywise matrix product, ego centrality is defined as $\text{ego}(i) = \Sigma_k \Sigma_{j>k} 1/[\mathbf{A}^2(i) \circ (\mathbf{I} - \mathbf{A}(i))]_{kj}$.

Local Fiedler vector centrality (LFVC) [13]: LFVC is a measure that characterizes vulnerability to node removals. A node with higher LFVC is more important for network connectivity structure. Let $\mathbf{y}$ (the Fiedler vector) denote the eigenvector associated with the second smallest eigenvalue $\mu(L)$ of the graph Laplacian matrix $\mathbf{L}$. LFVC is defined as $\text{LFVC}(i) = \Sigma_{j \in \mathcal{N}_i} (y_i - y_j)^2$. Although LFVC is a global centrality measure, it can be accurately approximated by local computations and message passing using the distributed power iteration method of [7] to compute the Fiedler vector $\mathbf{y}$.

The aforementioned centrality measures and their properties are summarized in Table 1.

Resilience of Western U.S. States Power Grid Topology to Centrality Attacks

A nodal centrality attack on a network incapacitates the nodes that have the highest centrality measures. The resilience of a network to centrality attacks is defined as the decrease in the size of the largest component that results from the attack. Throughout this article we adopt a greedy node removal strategy that sequentially removes the node with the highest centrality measure from the remaining largest component. The centrality measure is recalculated after node removals. It has been shown in [14] that greedy node removal strategies can be effective reducers of the largest component size compared to batch node removal strategies based on the same centrality measure. For general centrality measures there is no performance guarantee relating the greedy node removal strategy and the optimal batch removal strategy. However, using submodularity of the LFVC measure, it is proven in [13] that greedy node removal based on LFVC comes within at least $1 - 1/e$ of the performance of an optimal batch node removal strategy, where e is the Euler constant. Therefore, one might expect that greedy LFVC attacks are almost as effective as batch LFVC attacks in terms of severe impact on network connectivity.

We use the topology of the power grid of western U.S. states [8] to illustrate network vulnerability to different types of centrality attacks. The results are shown in Fig. 1. This network contains 4941 nodes and 6594 edges, where nodes represent power stations and edges represent power lines. More network topology information can be found in the supplementary file. One can see from Fig. 1 that an LFVC attack is capable of reducing the largest component size to roughly 54 percent of its original size by removing only 8 nodes from the network. On the other hand, betweenness and closeness attacks require å8 and 31 node removals, respectively, to achieve the same reduction. Equivalently, the LFVC attack

Input: number of rewires r, graph $G = (\mathcal{V}, \mathcal{E})$
Output: rewired graph $\widetilde{G} = (\mathcal{V}, \widetilde{\mathcal{E}})$
for $i = 1$ to r **do**
  Compute the second smallest eigenvector **y** of **L**
  Compute the largest eigenvector **z** of **L**
  Find $(i^*, j^*) = \arg\max_{(i,j)\notin\mathcal{E}} (y_i - y_j)^2$
  Find $(k^*, \ell^*) = \arg\max_{(k,\ell)\in\mathcal{E}} (z_k - z_\ell)^2$
  Edge addition stage: $\widetilde{\mathcal{E}} \leftarrow \mathcal{E} \cup (i^*, j^*)$
  Edge deletion stage: $\widetilde{\mathcal{E}} \leftarrow \widetilde{\mathcal{E}} / (k^*, \ell^*)$
  $G \leftarrow \widetilde{G}$
end for

Algorithm 1. Edge rewiring method.

requires removal of only 0.2 percent of the nodes in order to severely disrupt communications between nearly half of the nodes in the network. Furthermore, degree, eigen centrality, and ego centrality attacks fail to disrupt the network as significantly (less than 6 percent reduction in the largest component) even when 1 percent of the nodes are attacked. By inspecting the adjacency matrix **A** in [8], it is observed that the adjacency matrix has apparent blockwise structure where blocks are densely connected subgrids interconnected by relatively few inter-subgrid edges (see supplementary file). Since the high-degree nodes are not connected to those interconnected edges, and each subgrid is densely connected, greedy degree attacks do not result in severe connectivity loss. We conclude that LFVC attacks do significantly more damage than other types of centrality attacks. Therefore, LFVC is a more reliable measure of resilience of the network.

Preventive Approaches to Centrality Attacks

Here we discuss two preventive approaches to protect against centrality attacks: the edge addition method and the edge rewiring method.

The Edge Addition Method

Edge addition is perhaps the most intuitive method for enhancing resilience of network connectivity since it adds edges that are not already present in G. Let $\hat{\mathbf{L}}$ be the resulting graph Laplacian matrix after adding an edge $(i, j) \notin \mathcal{E}$ to G, and let **1** be a vector of all ones. Recalling the definition of the graph Laplacian matrix **L**, $\hat{\mathbf{L}} - \mathbf{L} = (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T$, where $\mathbf{e}_i$ is an all-zero vector except that its ith entry is equal to **1**. The term $(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^{\mathrm{T}}$ corresponds to the graph Laplacian matrix of the removed edge (i, j) alone. Since the algebraic connectivity $\mu(\mathbf{L})$ is the second smallest eigenvalue of **L**, and the smallest eigenvalue of **L** is 0 with associated eigenvector **1**, we have the representation $\mu(\mathbf{L}) = \min_{|\mathbf{x}|_2=1, \mathbf{x}^T\mathbf{1}=0} \mathbf{x}^T\mathbf{L}\mathbf{x}$ [9]. It is proved in [6] that

$$\mu(\hat{\mathbf{L}}) \geq \mu(\mathbf{L}) + c_1 \cdot (y_i - y_j)^2, \qquad (1)$$

where **y** is the eigenvector of $\mu(\mathbf{L})$ and $c_1 > 0$ is a positive constant.

Since algebraic connectivity is a lower bound on node connectivity and edge connectivity, it is proposed in [6] that one should iteratively add an edge that maximizes the quantity $(y_i - y_j)^2$ to the graph. For each iteration, the edge that maximizes $(y_i - y_j)^2$ maximizes the lower bound on the resulting algebraic connectivity, and therefore enhances network resilience to centrality attacks. The edge addition method will serve as the baseline for comparison to the proposed edge rewiring method.

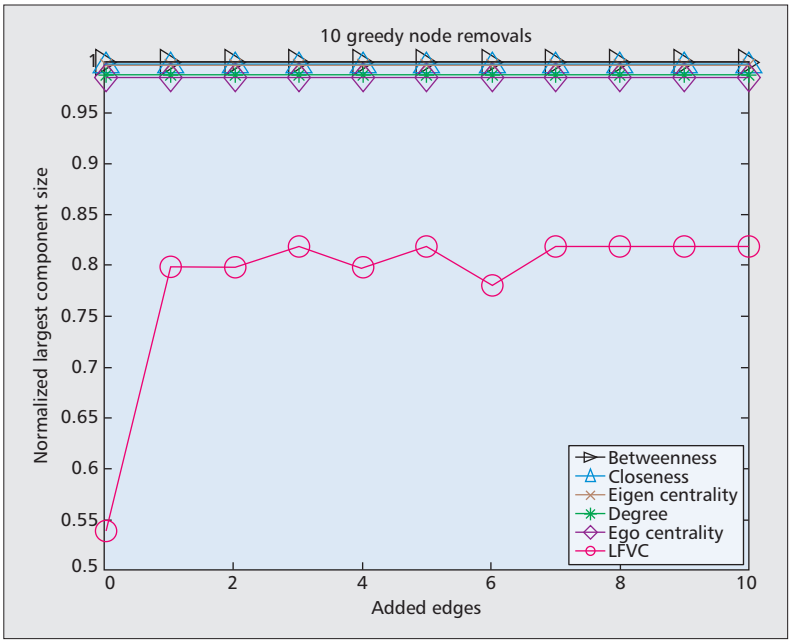


Figure 2. Network connectivity of the edge addition method when restricted to 10 greedy node removals on the power grid topology of western U.S. states [8]. The network connectivity can be enhanced from 54 to 80 percent under LFVC attacks by adding one edge.

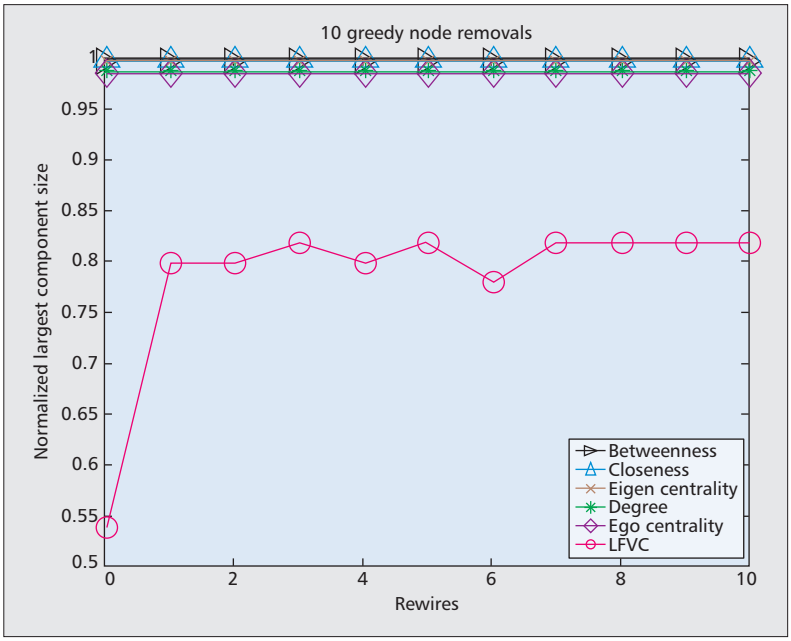


Figure 3. Network connectivity of the edge rewiring method when restricted to 10 greedy node removals on the power grid topology of western U.S. states [8]. The proposed edge rewiring method can perform as well as the edge addition method without introducing additional edges in the network.

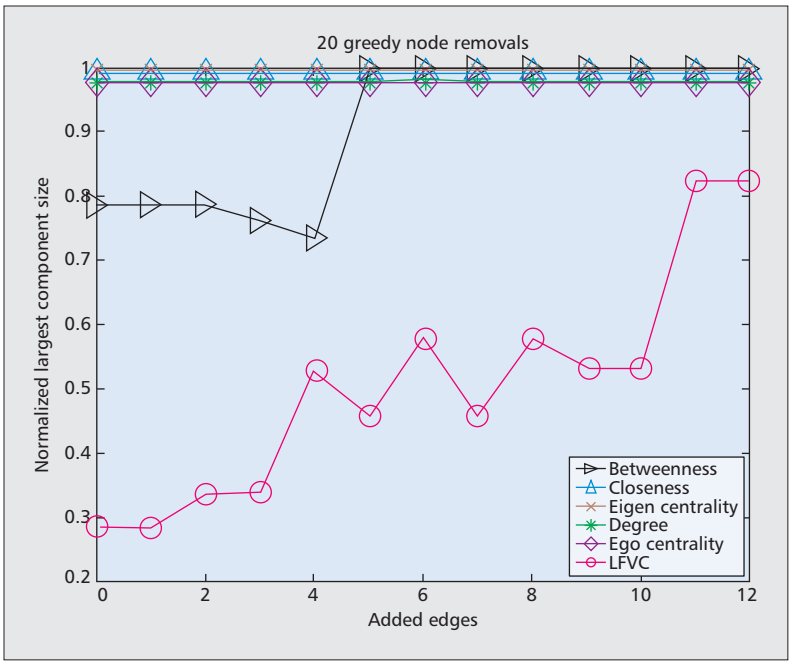


Figure 4. Network connectivity of the edge addition method when restricted to 20 greedy node removals on the power grid topology of western U.S. states [8]. Eleven additional edges are required to enhance the network connectivity from 29 to 82 percent.

The Edge Rewiring Method

Edge rewiring aims to rewire the edges in the graph in order to enhance the resilience of network connectivity to attacks. In particular, the edge rewiring method does not change the total number of edges in the graph. The proposed edge rewiring algorithm is summarized in Algorithm 1.

For each rewire, the edge rewiring method consists of two stages: an *edge addition* stage and an *edge deletion* stage. In the edge addition stage, similar to the edge addition method, an edge $(i, j) \notin \mathcal{E}$ that maximizes $(y_i - y_j)^2$ is selected to maximize the lower bound (Eq. 1) on the resulting algebraic connectivity. Let $\phi(\mathbf{L})$ denote the largest eigenvalue of $\mathbf{L}$, and let $\mathbf{z}$ denote the associated eigenvector of $\phi(\mathbf{L})$. In the edge deletion stage, an edge $(k, \ell) \in \mathcal{E}$ that maximizes $(z_k - z_\ell)^2$ is removed. The intuition is as follows. Let $\widetilde{\mathbf{L}}$ denote the graph Laplacian matrix after removing an edge from G. Since $\text{trace}(\mathbf{L}) - \text{trace}(\widetilde{\mathbf{L}}) = 2$ (i.e., 2 times the number of edge removals), and by Cauchy's eigenvalue interlacing property [15], $\phi(\mathbf{L}) \geq \phi(\widetilde{\mathbf{L}})$ and $\mu(\mathbf{L}) \geq \mu(\widetilde{\mathbf{L}})$, we have

$$\mu(\widetilde{\mathbf{L}}) \geq \mu(\mathbf{L}) + \phi(\mathbf{L}) - \phi(\widetilde{\mathbf{L}}) - 2. \qquad (2)$$

Consequently, for maximum effect, the edge rewiring algorithm should remove the edge that maximizes $\phi(\mathbf{L}) - \phi(\widetilde{\mathbf{L}})$ such that the lower bound on the resulting algebraic connectivity (Eq. 2) is maximized. By definition, $\phi(\mathbf{L}) = \max_{||\mathbf{x}||_2=1} \mathbf{x}^T \mathbf{L}\mathbf{x}$, and $\mathbf{L} - \widetilde{\mathbf{L}} = (\mathbf{e}_k - \mathbf{e}_\ell)(\mathbf{e}_k - \mathbf{e}_\ell)^T$ when the edge $(k, \ell) \in \mathcal{E}$ is removed. Therefore, computing $\mathbf{z}^T\widetilde{\mathbf{L}}\mathbf{z}$, we have $\phi(\mathbf{L}) - \phi(\widetilde{\mathbf{L}}) \leq (z_k - z_\ell)^2$. Moreover, by the eigenvector property that $\mathbf{z}$ is orthogonal to $\mathbf{1}$ (i.e., $\mathbf{z}^T\mathbf{1} = 0$), it is easy to verify that there is an edge $(k, \ell) \in \mathcal{E}$ and a constant $c_2 > 0$ such that $\phi(\mathbf{L}) - \phi(\widetilde{\mathbf{L}}) \geq c_2 \cdot (z_k - z_\ell)^2$.

Note that since the eigenvector $\mathbf{y}$ associated with $\mu(\mathbf{L})$ can be computed in a distributed manner [7], the eigenvector $\mathbf{z}$ associated with $\phi(\mathbf{L})$ can also be obtained using distributed local computations and message passing.

Performance Evaluation

In this section, we evaluate the effectiveness of the edge addition and edge rewiring methods in protecting the power grid topology [8] from centrality attacks. When 10 nodes are removed from the network by LFVC attacks, Fig. 1 shows that the network connectivity is reduced to 54 percent. In contrast, under other types of centrality attacks there is almost no loss in connectivity when 10 nodes are removed. Figure 2 illustrates the effect of edge addition as a preventive approach against centrality attacks. It is observed that by adding one edge, the network connectivity can be increased from 54 to 80 percent under LFVC attack. Figure 3 illustrates the proposed edge rewiring method. Similar to the edge addition method, one edge rewire is capable of enhancing the network connectivity from 54 to 80 percent. Thus, using the edge rewiring method with only one edge rewire can protect the network as well as the edge addition method even though the latter introduces additional edges in the network.

When 20 nodes are removed from the network, as shown in Fig. 4, 11 edge additions are required to increase network connectivity from 29 to 82 percent. In comparison, as shown in Fig. 5, the proposed edge rewiring method requires only 12 edge rewires to achieve the same performance, which means that we only need to rewire fewer than 0.4 percent of the edges to make it resilient to centrality attacks. This performance advantage is explainable since, for the same number of edge additions or rewiring actions, edge rewiring changes twice as many edges in the network as edge addition. A second illustrative example for a European Internet backbone network is discussed in the supplementary file.

Conclusion and Future Work

This article investigates network resilience to centrality attacks, proposes a centrality measure for assessing resilience, and studies two preventive approaches for protecting networks against such attacks. The results on the power grid of western U.S. states show that the network is particularly vulnerable to LFVC attacks, and that the edge rewiring method can significantly improve network resilience with only a few edge rewires. Useful areas for future work are:

- Extension to time-varying topologies
- Extension to topologies with weighted edges
- Application to social networks

Acknowledgment

This work has been partially supported by the Army Research Office (ARO), grant number W911NF-12-1-0443.

References

[1] R. Albert, H. Jeong, and A.-L. Barabási, "Error and Attack Tolerance of Complex Networks," *Nature*, vol. 406, no. 6794, July 2000, pp. 378–82.

[2] S. Xiao, G. Xiao, and T. H. Cheng, "Tolerance of Intentional Attacks in Complex Communication Networks," *IEEE Commun. Mag.*, vol. 45, no. 1, Feb. 2008, pp. 146–52.

[3] P.-Y. Chen, S.-M. Cheng, and K.-C. Chen, "Smart Attacks in Smart Grid Communication Networks," *IEEE Commun. Mag.*, vol. 50, no. 8, Aug. 2012, pp. 24–29.

[4] P.-Y. Chen, S.-M. Cheng, and K.-C. Chen, "Information Fusion to Defend Intentional Attack in Internet of Things," *IEEE Internet Things J.*, vol. 1, no. 4, Aug. 2014, pp. 337–48.

[5] P.-Y. Chen and A. Hero, "Node Removal Vulnerability of the Largest Component of a Network," *Proc. IEEE GlobalSIP*, 2013, pp. 587–90.

[6] A. Ghosh and S. Boyd, "Growing Well-Connected Graphs," *Proc. IEEE Conf. Decision and Control*, 2006, pp. 6605–11.

[7] A. Bertrand and M. Moonen, "Distributed Computation of the Fiedler Vector with Application to Topology Inference in Ad Hoc Networks," *Signal Processing*, vol. 93, no. 5, 2013, pp. 1106–17.

[8] D. J. Watts and S. H. Strogatz, "Collective Dynamics of "Small-World" Networks," *Nature*, vol. 393, no. 6684, June 1998, pp. 440–42; http://www-personal.umich.edu/mejn/netdata

[9] M. Fiedler, "Algebraic Connectivity of Graphs," *Czechoslovak Math. J.*, vol. 23, no. 98, 1973, pp. 298–305.

[10] L. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, 1977, pp. 35–41.

[11] G. Sabidussi, "The Centrality Index of a Graph," *Psychometrika*, vol. 31, no. 4, 1966, pp. 581–603.

[12] M. Everett and S. P. Borgatti, "Ego Network Betweenness," *Social Networks*, vol. 27, no. 1, 2005, pp. 31–38.

[13] P.-Y. Chen and A. Hero, "Local Fiedler Vector Centrality for Detection of Deep and Overlapping Communities in Networks," *Proc. IEEE ICASSP*, 2014, pp. 1120–24.

[14] P. Holme *et al.*, "Attack Vulnerability of Complex Networks," *Phys. Rev. E*, vol. 65, May 2002, p. 056109.

[15] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, 1990.

Biographies

Pin-Yu Chen [S'10] received a B.S. degree in electrical engineering and computer science (undergraduate honors program) from National Chiao Tung University, Taiwan, in 2009, an M.S. degree in communication engineering from National Taiwan University in 2011, and is currently working toward a Ph.D. degree in electrical engineering and computer science at the University of Michigan, Ann Arbor. He is a member of the Tau Beta Pi Honor Society and the Phi Kappa Phi Honor Society, and was the recipient of the Chia-Lun Lo Fellowship. He was also the recipient of the IEEE GLOBECOM 2010 GOLD Best Paper Award. His research interests include network science, interdisciplinary network analysis, and their applications to communication systems.

Alfred O. Hero III [F] received his B.S. (summa cum laude) from Boston University (1980) and Ph.D from Princeton University (1984), both in electrical engineering. Since 1984 he has been with the University of Michigan, Ann Arbor, where he is the R. Jamison and Betty Williams Professor of Engineering. His primary appointment is in the Department of Electrical Engineering and Computer Science, and he also has appointments, by courtesy, in the Departments of Biomedical Engineering and Statistics. From 2008 to 2013 he held the Digiteo Chaire d'Excellence at the Ecole Superieure d'Electricite, Gif-sur-Yvette, France. Several of his research articles have recieved best paper awards. He was awarded the University of Michigan Distinguished Faculty Achievement Award (2011). He received the IEEE Signal Processing Society Meritorious Service Award (1998), the IEEE Third Millenium Medal (2000), and the IEEE Signal Processing Society Technical Achievement Award (2014). He was President of the IEEE Signal Processing Society (2006–2008) and on the Board of Directors of the IEEE (2009–2011) where he served as Director of Division IX (Signals and Applications). His recent research interests are in statistical signal processing, machine learning, and the analysis of high dimensional spatio-temporal data. Of particular interest are applications to networks, including social networks, multimodal sensing and tracking, database indexing and retrieval, imaging, and genomic signal processing.

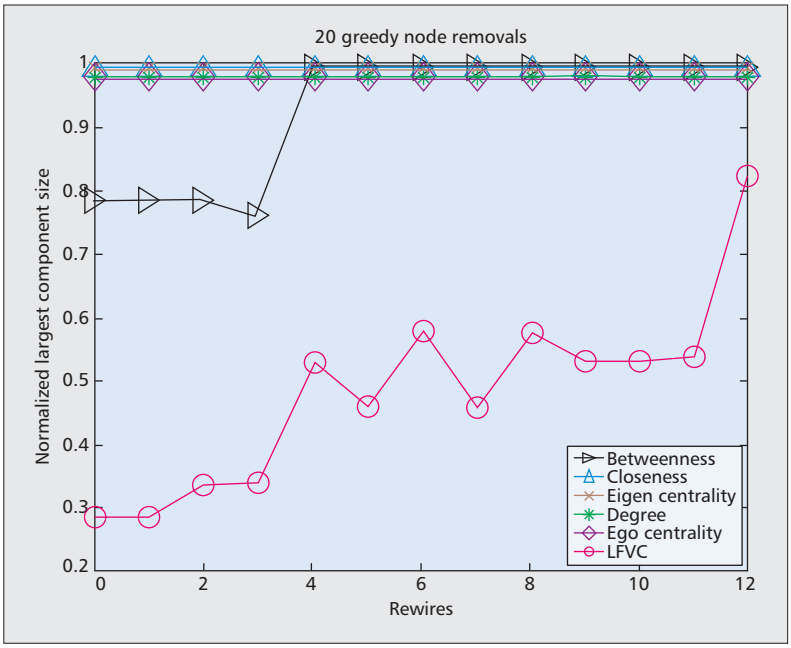


Figure 5. Network connectivity of the edge rewiring method when restricted to 20 greedy node removals on the power grid topology of western U.S. states [8]. The proposed edge rewiring method requires only 12 edge rewires to achieve the same performance as in Fig. 4, which means that we only need to rewire fewer than 0.4 percent of edges to make it resilient to centrality attacks.

ACCEPTED FROM OPEN CALL

Physics-Inspired Methods for Networking and Communications

David Saad, Chi Ho Yeung, Georgios Rodolakis, Dimitris Syrivelis, Iordanis Koutsopoulos, Leandros Tassiulas, Rüdiger Urbanke, Paolo Giaccone, and Emilio Leonardi

ABSTRACT

Advances in statistical physics relating to our understanding of large-scale complex systems have recently been successfully applied in the context of communication networks. Statistical mechanics methods can be used to decompose global system behavior into simple local interactions. Thus, large-scale problems can be solved or approximated in a distributed manner with iterative lightweight local messaging. This survey discusses how statistical physics methodology can provide efficient solutions to hard network problems that are intractable by classical methods. We highlight three typical examples in the realm of networking and communications. In each case we show how a fundamental idea of statistical physics helps solve the problem in an efficient manner. In particular, we discuss how to perform multicast scheduling with message passing methods, how to improve coding using the crystallization process, and how to compute optimal routing by representing routes as interacting polymers.

INTRODUCTION

Large communication networks, such as the Internet, are very complex distributed systems, composed of millions of architectural elements, interacting according to complex and unpredictable patterns. Despite the fact that the Internet is man made and its internal element architectures and protocols are well known and standardized, its overall behavior is still largely unclear. On this regard we quote Eric Schmidt, Google-CEO: "The Internet is the first thing that humanity has built that humanity doesn't understand, the largest experiment in anarchy that we have ever had."

For these reasons, analyzing, optimizing, and predicting the performance of large scale communication networks raises several challenges. Traditional approaches, such as Markov chains, control theory, queuing theory, Monte-Carlo simulations, which have been largely employed in network design and protocol analysis, permit a very detailed microscopic description of the dynamics of some network elements, but fail to provide scalable tools for the comprehension of emerging dominant macroscopic dynamics.

Recently, techniques borrowed from statistical physics have been successfully applied to represent emerging macroscopic phenomena in computer networks: Mean field approaches have been employed to analyze the dynamics of congestion control algorithms (such as TCP) in high capacity networks where thousands of users share the limited bandwidths [1]. In the context of wireless networks, percolation theory results have been applied for capacity analysis of large-scale networks [2]. A parallelism with interacting Fermionic particle systems has been invoked to explain phase transition phenomena observable in large scale dense Wi-Fi networks [3]. A new optimal-performing MAC protocol, inspired by Glauber dynamics over graphs, has been proposed as a viable alternative to the current Wi-Fi scheme [4]. Message passage algorithms have been employed to solve specific packet scheduling problems in a scalable and distributed fashion [5]. Other methods adopted from statistical physics to address problems in communication and networking have been reviewed in [6].

Despite the rich body of work on employing statistical physical concepts in networks, a systematic statistical mechanics-inspired networking theory is still missing. This survey is a first step, based on the EU-FP7 project STAMINA, to fill this gap in part, by providing a general introduction to statistical physics methods that can be broadly employed for performance analysis of communication networks as well as network design and optimization. The main advantage offered by statistical physics methods is in the range of tools developed to address large scale problems of a non-linear nature, through the study of *typical* case behavior and relying on approximation techniques that are well established within the physics community. The disadvantage of these methods is that many of them are non-rigorous and that some tend to fail in small-scale systems.

The methods suggested throughout the article all stem from the statistical physics of disordered systems and have been adapted to the various problems depending on the suitability of a particular tool to the problem at hand. In particular, we present a description of a suitable statistical physics framework and, via indicative application examples, the potential of three

David Saad and Chi Ho Yeung are with Aston University, UK.

Georgios Rodolakis, Dimitris Syrivelis, and Iordanis Koutsopoulos are with Centre for Research and Technology Hellas (CERTH).

Leandros Tassiulas is with Yale University.

Rüdiger Urbanke is with Ecole Fédérale Polytechnique de Lausanne (EPFL).

Paolo Giaccone and Emilio Leonardi are with Politecnico di Torino.

0163-6804/14/$25.00 © 2014 IEEE

classes of techniques: multicast scheduling and message passing; reliable transmission; and crystallization, routing and interacting polymers.

Statistical Physics and Communication Problems

So what can physics contribute to the quest for improved and more principled solutions for increasingly complex and difficult communication problems?

Among the main characteristics of statistical physics methods is that they deal with large-scale systems where interactions between system variables are non-linear and exhibit macroscopic emergent behavior, which is a characteristic of complex systems. Sophisticated techniques have been developed in this area to specifically address systems of this type and can be exploited to gain insight and develop optimization algorithms for real systems of similar characteristics.

The type of physical systems most suitable to represent optimization and constrained problems are termed disordered systems, where the disorder may refer to the underlying topology, that is, the specific choice of edges in a graph, or strength of interaction between system constituents; these are assumed to be drawn from some distribution. Statistical mechanics methods for disordered systems in equilibrium revolve mostly around the concept of free energy that relates to the probability of the system to be in a given state. Calculating the averaged *free energy* over instances of the disorder which appears implicitly in the form of interaction between the various variables is difficult and is the focus of the methods mentioned below.

Among the main methods adopted from the physics of disordered systems for the study of complex systems are the cavity and replica methods [7]. They facilitate the calculation of averages over all possible system instances in order to find the macroscopic system behavior, for instance over all networks of N vertices and given degree connectivity. While these methods focus on providing insight into the macroscopic behavior of the systems investigated, they usually also give rise to algorithms, such as belief propagation, that allow for the inference of variable values in specific instances.

Multicast Scheduling and Message Passing

Application: Scheduling

To keep up with the growing demand for communication resources in a cost-effective manner, optimal resource management becomes imperative, thus giving rise to hard optimization problems. One such example is scheduling, where a global objective, such as total throughput or energy-consumption, has to be optimized while satisfying local constraints, such as conflicts between simultaneous wireless transmissions or simultaneous packet transmissions toward the same port interface in routers.

In particular, scheduling multicast traffic in Input Queue (IQ) switches requires solving a hard combinatorial optimization problem (the scheduling decision for each packet) in a very short time, comparable with the packet transmission time at the port interfaces. Consider the IQ switch illustrated in Fig. 1a. Based on the state of occupancy of the queues, the scheduler, must select a set of non-conflicting packets in order to maximize the total throughput. For multicast traffic, one logical queue is present for each possible input port and fanout set (i.e. a subset of output port destinations), thus making the scheduler's task highly complex.

This example constitutes an indicative testcase of the potential of the broader applicability of message passing techniques in communications. In this context, challenging optimization problems often require the design of algorithms that are both easy to implement and efficient in terms of performance.

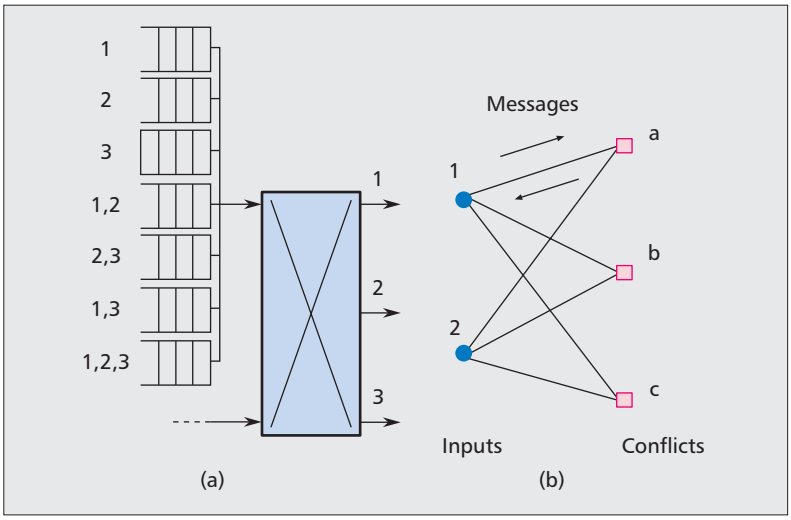


Figure 1. a) An IQ switch with 2 inputs and 3 outputs. Each logical queue is tagged with the corresponding fanout set; b) a toy example of a bipartite graphical model of the switch with blue nodes corresponding to inputs and red nodes to interaction terms (conflicting multicast packet transmissions).

Physics: Cavity Method and Message-Passing

The optimization problem is mapped onto a statistical physics problem with a probability distribution over possible configurations, and the computation of the optimal solution is reduced to identifying the minimum-energy configurations at which the probability distribution concentrates. In order to find a configuration of low energy, one can use statistical physics techniques such as the cavity method. The latter is based on calculating the influence of neighboring nodes in the absence of the node in question, which leads to a set of coupled equations that can be solved iteratively, similar to density evolution, the macroscopic equivalent of belief propagation algorithms.

A convenient *graphical model* of the underlying system is a bipartite graph, with one set of nodes representing the system components (e.g. input ports of the IQ switch), while the other set consists of local interaction terms (such as con-

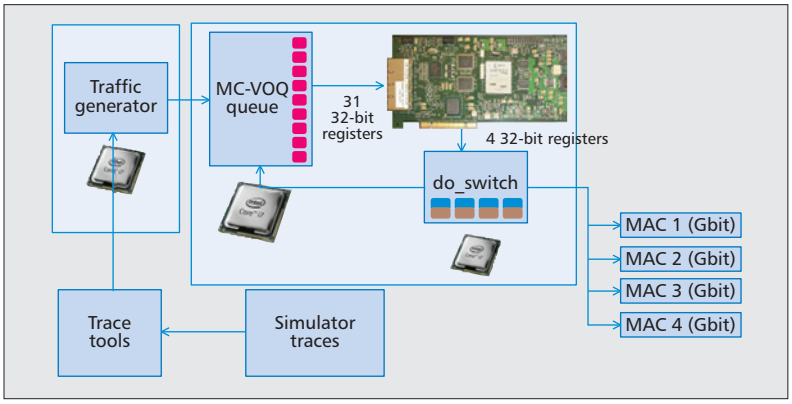


Figure 2. A hardware/software co-design schematic of BP scheduling implemented on NetFPGA.

flicts in simultaneous packet transmissions). In Fig. 1b we depict a toy example with two input ports and three interaction terms, modeling potential conflicts in queued packets due to intersecting multicast destination sets.

Each interaction node a sends a message $m(a \to i)$ to each neighbor node which contains a's belief about the state of node i. This belief is essentially a conditional probability estimated by a, based on all messages it receives from nodes other than i. Similarly, each node i computes its own belief and sends message $m(i \to a)$. These messages are computed separately by each node, they are iteratively updated and propagate through the system graph. This low-overhead iterative procedure usually leads to a low-energy configuration; in the case where the graph is a tree, it can be shown to converge to the minimal-energy configuration. Thus, centralized problems can be solved in a distributed manner with iterative lightweight local messaging.

Recent success stories in discrete computational challenges attest to the strong potential of these techniques, including Shannon-capacity approaching codes and NP-hard problems like K-satisfiability [7], leading to more recent applications in wireless scheduling problems [5].

How Physics Helps in the Application

We present a specific application of the statistical physics methodology in the novel design and hardware implementation of a multicast switch.

The throughput-optimal scheduling policy for multicast traffic allows for "fanout-splitting", that is, a packet can be sent to just a subset of its destination ports, leaving some residual destinations for future transmissions. The main idea of the optimal policy is to serve at higher priority packets that are stored in large queues and that are possibly re-enqueued into smaller ones. In order to solve the resulting problem, one can resort to a Belief Propagation (BP) algorithm [8].

Messages are exchanged between each input and output ports, and are updated iteratively by each port concurrently. The scheduler chooses the packets to be transferred, based on the final *beliefs*, that is, local estimates at each input port of the throughput that can be achieved by choosing specific transmission fanout sets. Due to the "densely connected" constraints that prevent conflicting packets, there are several cases in which the message update phase does not converge. Such a difficulty can be overcome using BP with a fixed number of iterations, in conjunction with a centralized algorithm, which at each step chooses the transmission fanout set with the maximum belief.

The BP approach outperforms other greedy algorithms (such as longest queue first) in simulations, with a gain between 6 percent and 48 percent under uniform traffic, and between 5 percent and 10 percent under worst-case concentrated traffic [8]. Interestingly, only a very small number of message update iterations is necessary to achieve this performance (as low as 1 or 2).

We have implemented the belief propagation scheduling algorithm as a hardware accelerator on the netFPGA platform [9]. The implementation consists of a software communication interface and a hardware scheduler state machine for a 4×4 switch, integrated in the emulation framework illustrated in Fig. 2. The measured duration of the scheduling algorithm execution is 3.77μs, 5.44μs, and 7.12μ for 0, 1 and 2 BP iterations, respectively, that is, a 2.65× improvement over the performance if implemented only in software. In terms of energy efficiency, the power consumption of the hardware switch during the scheduling algorithm execution is 26.1W, versus the platform's idle power consumption of 23.8W.

This successful implementation demonstrates the feasibility and potential of the message passing approach in practical networking problems, with stringent requirements on efficiency and lightweight operation.

Reliable Transmission, Crystallization and the Nucleation Phenomenon

Application: Reliable Transmission of Information

The reliable transmission of information is at the heart of any communication system. Whether it is noise due to thermal effects or packet losses due to buffer overflows, there are many physical phenomena and processes that lead to a loss or degradation of transmitted information. Error-correcting coding is the means of ensuring that, despite all these adverse effects, the end-to-end communication link is reliable.

Traditionally, error-correcting codes were based on algebraic notions of ever increasing sophistication, and codes were designed to maximize the Hamming distance between codewords, that is, the minimum number of positions in which two distinct codewords differ. But in the last 20 years, codes based on sparse graphs and message-passing schemes have fundamentally changed the way codes are designed and have gradually replaced traditional schemes.

Sparse graph codes are based on bipartite graphs, as shown in Fig. 3 where the length of the code is $N = 7$.

Each round node on the left represents a bit, and each square node on the right represents a constraint. In particular, these constraints represent linear equations that have to be fulfilled. The code is the set of all binary sequences of

length N that fulfill all these constraints. To transmit information we pick a codeword, transmit it, and use the redundancy that is inherent in the code to recover the transmitted word from the received information.

The term *sparse* graph code indicates that the blocklength N is typically in the thousands, but the degrees of both variables and checks are taken from a finite set and do not depend on N. Hence the number of edges is of order N and not N^2 and is therefore "sparse." The important point about such sparse graph codes is that their decoding is accomplished by the belief propagation algorithm; that is, given a noisy version of the transmitted codeword, messages representing the current "beliefs" about the various bits are exchanged along the edges until these messages (hopefully) converge to the correct values. Such a decoder is inherently of low complexity and can conveniently be implemented in hardware.

We use such codes daily, since they are part of state-of-the art cell phones, Wi-Fi modems, optical transmission schemes, and hard drives. Despite intensive research on sparse graph codes and their wide deployment, there is still room for improvement. In particular, it is difficult to design a coding scheme that allows for reliable transmission close to capacity and that has very low error probabilities as required, for example, for storage applications (hard disks) or in the backbone of the Internet (optical communications). This is due to the fact that in order to achieve transmission close to capacity, typically a large number of variable nodes of small degree are necessary but also cause relatively high "error floors." Moreover, such codes are usually not universal, that is, codes that are designed and reliable for one channel might not allow reliable transmission over another, even if it has equal capacity. Quite recently, an interesting physical phenomenon has been shown to be useful in overcoming these two difficulties.

Physics: Crystallization

Crystallization is the process that describes how solid crystals form from a liquid solution. This solution is typically a *meta-stable* state, which does not correspond to the lowest-energy configuration and which therefore eventually goes into the stable crystalline form. Nevertheless, the meta-stable state can persist on long time scales and a *nucleus*, that is a seed, is needed to get the crystallization process started.

Nucleation and the crystallization process can take on several forms, and we are all familiar with them in several disguises:

- Reusable heat packs typically contain sodium acetate enclosed in a suitable container. When heated, sodium acetate takes on a liquid form that is a meta-stable state that is stable over long periods of time. The nucleation process takes on the form of bending a small metal disc that is contained in the heat pack. This starts the crystallization process and thereby releases heat.
- Water can be brought into a supercooled state in that it is still in liquid form even though it is considerably below the freezing point, as long as it is cooled in a clean container. The nucleation process starts by, for example, shaking the container violently or by adding a little seed. A quick search on YouTube for "supercooled water" shows many instances how the crystallization process starts suddenly through a suitable nucleation process. Other examples are cloud or hail formation.

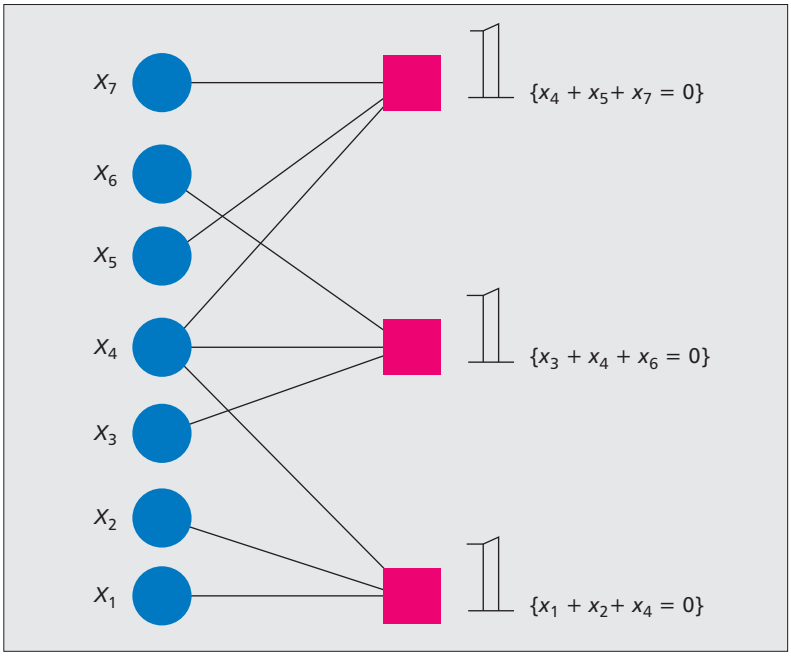


Figure 3. Bipartite graph representing a sparse graph code of length $N = 7$. The code consists of the set of binary words of length 7 that fulfill the three linear constraints.

How Physics Helps in the Application

The nucleation phenomenon can be exploited to build codes that are provably capacity-achieving under message-passing decoders and universal for large classes of channels. Figure 4 shows an incarnation of the basic idea.

Rather than using an unstructured graphical model, we "spatially couple" a number of such graphical models along a chain in such a way that neighboring models interact, but that the local degree structure of each model stays unchanged. In addition, we properly "terminate" the chain in such a way that the problem is made easier at the boundary.

Applying the basic belief-propagation algorithm on such a code, an interesting phenomenon occurs. The code can be successfully decoded up to a higher noise value than what is possible for the underlying "component" code. In fact, codes constructed in such a way can be decoded up to the maximum a-posteriori threshold of the underlying code. This is the highest threshold achievable under any decoding algorithm, no matter how complex. The decoding happens along a "wave." At the boundary, due to the special termination, decoding is easier and bits are decoded first. Once the bits at the boundary have been decoded, the "interface" between the decoded and uncoded parts moves inward, and this decoding wave advances at a constant speed.

Mathematically and physically there is a

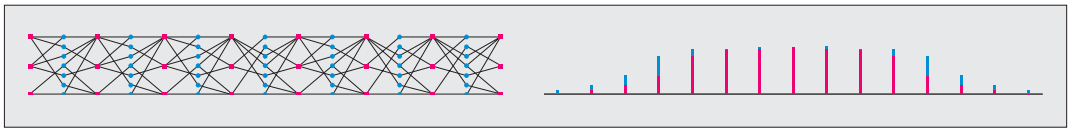


Figure 4. Left: A spatially coupled code. Several "component" codes are placed along a line are neighboring components are "coupled" that is, they share edges. At the boundary there are locally more check nodes than in the interior. This makes decoding easier and acts as a seed. Right: The "decoding wave." The horizontal axis corresponds to the spatial dimension of the code. The vertical dimension shows that local error probability at a certain point in the decoding process. As the number of iterations increases, the error probability decreases taking the shape of a "wave" that moves at constant speed from the boundary towards the middle.

direct analogy to nucleation and the crystallization process. The special termination at the boundary acts as the nucleus and gets the decoding started. Without the boundary the belief-propagation decoder, which is in general suboptimal, is not strong enough to drive the system into the lowest-energy state, which corresponds to the correct, that is, transmitted codeword, but instead ends up in a meta-stable state. But once decoding has started at the boundary, the process continues like in the crystallization process, and the whole system moves at a constant speed toward the lowest-energy configuration. Codes constructed based on this principle can be designed to work arbitrarily close to the capacity of the channel with high reliability and are inherently universal, that is, they work well over whole classes of channels and do not have to be tuned to a particular application [10, 11].

This phenomenon can also be exploited in other areas. For example, it has been used in compressive sensing [12] and to analyze constraint satisfaction problems [13].

Routing and Polymers

Application: Routing

Finding optimal routes, given some measure of optimality, is a difficult task with implications for a large number of application domains, from water distribution networks and VLSI design to journey planners. Clearly, it has many applications in the area of communication networks, ranging from sensor and optical networks to peer-to-peer and wireless communication.

Optimal routing is also a very hard computational problem, being non-localized with non-linear interactions at vertices (routers) and/or edges (communication lines); therefore, most existing routing algorithms are based on localized selfish decisions and rely on (mostly non-adaptive) routing tables to identify the shortest weighted path to the destination regardless of the individual decisions made, for example, the celebrated Dijkstra algorithm. Dynamic routing protocols do exist, but they are mostly heuristic and insensitive to other individual routing decisions that dynamically constitute the traffic.

To optimize the use of resources, a more global approach is required that takes into account all individual routing decisions and makes efficient use of the over-stretched network infrastructure. The cost to be minimized is defined according to the task at hand. For instance, in many cases one would like to suppress congestion in order to avoid bottlenecks by minimizing overlaps between routes, either at vertices or at edges, possibly attributing weights to vertices and/or edges to reflect preference, capacity, or delays. At the other end of the spectrum one may employ an objective function that aims to decrease the number of active vertices by consolidating paths to reduce infrastructure demands or energy consumption. This is particularly relevant in the context of communication networks, as the Internet can consume up to 4 percent of the electricity generated at peak times. These objective functions are typically non-linear and represent interactions between non-localized objects — multiple routes.

Physics: Polymers

Although the techniques developed for the study of disordered systems, such as the cavity method (message passing), have the potential to model interactions between simple system constituents such as network nodes, they may be difficult to apply when interactions involve more complicated objects such as routes. In the case of routing, additional techniques should be employed in order to verify that routes are contiguous, leading from sources to destinations. Techniques developed in the study of polymers are ideally suited for these tasks.

A polymer can be viewed as a chain of molecules connected to one another in a manner where the end of one segment is the beginning of another. Routes on a network can be viewed as polymers placed on a graph such that segments correspond to edges and interact at vertices, as shown in Fig. 5. The aim is to choose the shortest routes (polymers) while incorporating an interaction between them on vertices or edges: repulsion, in case one wants to minimize congestion and make traffic as uniform as possible; and attraction, when one wants to consolidate routes and reduce the number of active vertices and/or edges.

How Physics Helps in the Application

We studied two routing scenarios using statistical physics methods:

- The case in which randomly selected nodes on a given graph communicate with specific preassigned router(s) while minimizing congestion or maximizing route consolidation [14]. This study revealed the macroscopic behavior in terms of cost and average path length as the number of communication sources increases and provided a distributed algorithm to find optimal solutions in specific instances.

• A scenario whereby we analyze macroscopic properties of generic path-optimization problems between arbitrarily selected communicating pairs; we also derive a simple, principled, generic and distributive routing algorithm, capable of considering simultaneously all individual path choices.

While the method is suitable for many networking and communication problems, we choose a more graphical and real-life problem to demonstrate the efficacy of the new algorithm: routing passengers on the London underground network based on real Oyster-card data. Figure 6a shows how congestion is reduced by the algorithm when the objective function chosen aims to repel routes, as reflected in the fairly uniform traffic distribution even in the central region; the cost obtained by our algorithm is 20.5 percent smaller than that of the shortest path configuration obtained by the Dijkstra algorithm and slightly better than other state-of-the-art algorithms [15], with only a slight increase in average path length by 5.8 percent. In contrast, Fig. 6b shows how paths for the same passenger set are consolidated at major routes and stations, when a cost aimed at consolidating routes is chosen. This scenario may be relevant at times when the service is reduced for some reason, for instance during a strike or at off-peak hours to decrease costs. Due to the concave nature of the cost, there are no efficient competitive algorithms for carrying out these tasks.

These methods have direct relevance to various communication problems, such as node-disjoint routing, a hard-computational problem that is essential to prevent blocking in optical networks [16]. The insight gained by employing methods of statistical mechanics to routing problems and the efficient algorithms derived are of great potential and will help provide more efficient and scalable individualized routing algorithms.

The main challenges one needs to address for making these routing techniques more applicable are: a) to accommodate temporal interaction between routes, that is, taking into account the time it takes to arrive at nodes/edges and limiting the interaction between routes to concurrent traffic; b) devising approximation techniques in the case of node/edge-disjoint multi-colored routing in the presence of a large number of different colors, such as in the case of realistic multi-wavelength optical networks.

Conclusion and Future Perspectives

Physics-inspired methods can lead to breakthroughs when applied to communications and networking tasks. In this article we listed some instances of such problems. The presentation spans a range of examples, from aspects of theoretical modeling using physics-inspired methods to how such methods find their way into systems that are amenable to implementation.

The range of possible applications of physics-inspired methods is in no way limited to the three applications we discussed here. In concluding, we discuss three challenging research directions.

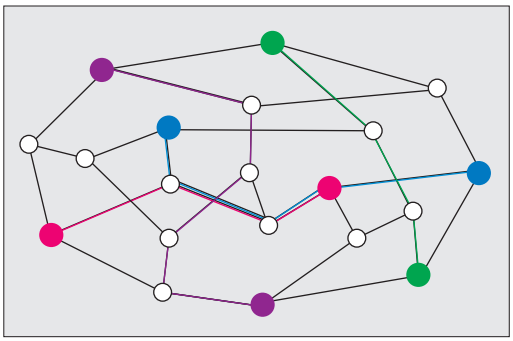


Figure 5. Routes as polymers on a regular random network of degree 3. Full circles represent start and end points of communication paths, each one presented in a different color, while intermediate empty nodes may constitute part of the path(s). Interaction between polymers may take place on vertices and/or on edges.

Non-Equilibrium Methods — Most existing physics-inspired methods have been developed for understanding homogeneous equilibrium systems, while most real communication systems are heterogeneous, hierarchical, and inherently not in equilibrium. Adapting non-equilibrium methods from statistical physics to manage and optimize dynamical and hierarchical communication systems is a significant challenge. This new approach may also reveal hidden network properties and shed light on system design choices.

Big Data — The amount of data generated in recent years is unprecedented and, more often than not, the rate of generation exceeds the rate at which they can be absorbed and understood, by humans or machines. Often data streams should be processed online and low-complexity schemes are needed, to create meaningful information out of the deluge of data. Being equilibrium-based methods, current message passing methods are limited in their ability to deal with data streams, and novel dynamical inference methods should be devised to address this challenge.

Capacity Crunch of Communication Networks — The ever-increasing demand for information, both due to the increasing number of mobile networks and the powerful and information-hungry networked devices, bring communication networks close to capacity. Smartphones, tablets, and watches with multiple embedded sensors capture data from their surroundings and transmit it; incoming mobile video traffic is also increasing at an exponential rate. Clearly this introduces unprecedented requirements to communication network operators and challenges the provisioning of end-user Quality of Experience. Novel schemes are needed, in multi-dimensional resource configurations, to optimize content transmission to the end-user. For example, intelligent caching of content at various locations or even at neighboring devices is needed in conjunction with joint optimization across resource domains like energy, computational

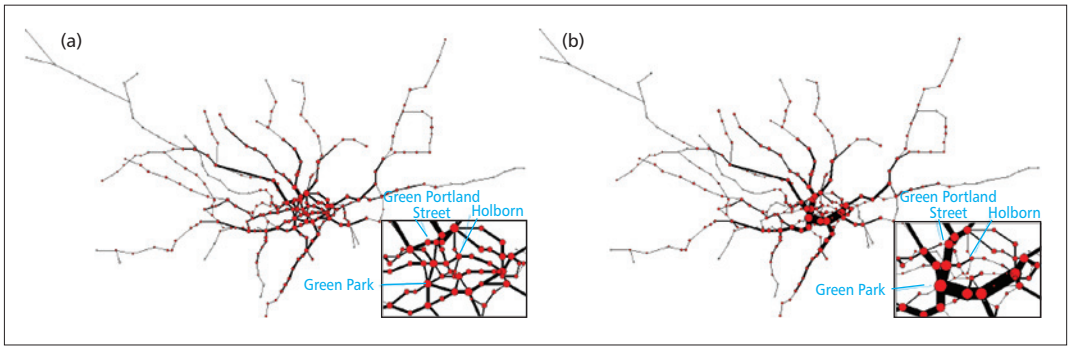


Figure 6. Optimized traffic on the London subway network. A total of 218 real passenger source-destination pairs are optimized, corresponding to 5 percent of data recorded by the Oyster card system between 8:30am – 8:31am on one Wednesday in November 2009. The network consists of 275 stations. The corresponding costs per node are chosen to a) repel and b) consolidate routes. Red nodes correspond to stations. The size of each node and the thickness of each edge are proportional to traffic through them.

capacity, and control, the latter being reflected in the choice of transmission rate and power, access point association, and so on. Physics-inspired techniques could be crucial in optimizing performance with multiple objectives and constraints.

Acknowledgments

This work was supported by the EU FP7-265496 project STAMINA (http://stamina-ict.eu).

References

[1] F. Baccelli *et al.*, "A Mean-Field Analysis of Short Lived Interacting TCP Flows," *Proc. Joint International Conf. Measurement and Modeling of Computer Systems*, SIGMETRICS '04/Performance '04, New York, NY, USA, 2004, ACM, pp. 343–54.
[2] M. Franceschetti *et al.*, "Closing the Gap in the Capacity of Wireless Networks via Percolation Theory," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1009–18, March 2007.
[3] M. Durvy, O. Dousse, and P. Thiran, "On the Fairness of Large CSMA Networks," *IEEE JSAC*, vol. 27. no. 7, Sept. 2009, pp. 1093–1104.
[4] L. Jiang *et al.*, "Fast Mixing of Parallel Glauber Dynamics and Low-Delay CSMA Scheduling," *Proc. IEEE INFOCOM*, 2011, April 2011, pp. 371–75.
[5] S. Sanghavi, D. Shah, and A. S. Willsky, "Message Passing for Maximum Weight Independent Set," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, 2009, pp. 4822–34.
[6] C. H. Yeung and D. Saad, "Networking — A Statistical Physics Perspective," *J. of Physics A: Mathematical and Theoretical*, vol. 46, no. 10, 2013, p. 103001.
[7] M. Mezard and A. Montanari, *Information, Physics, and Computation*, Oxford University Press, USA, 2009.
[8] P. Giaccone and M. Pretti, "A Belief-Propagation Approach for Multicast Scheduling in Input-Queued Switches," *Proc. Workshop on Networking Accross Disciplines: Communication Networks, Complex Systems and Statistical Physics* (NETSTAT), 2013, pp. 1403–08.
[9] D. Syrivelis *et al.*, "On Emulating Hardware/Software Co-Designed Control Algorithms for Packet Switches," *Emutools*, 2014.
[10] A. J. Felström and K. S. Zigangirov, "Time-Varying Periodic Convolutional Codes with Low-Density Parity-Check Matrix," vol. 45, no. 5, Sept. 1999, pp. 2181–90.
[11] S. Kudekar, T. Richardson, and R. L. Urbanke, "Spatially Coupled Ensembles Universally Achieve Capacity Under Belief Propagation," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, Dec 2013, pp. 7761–7813.
[12] F. Krzakala *et al.*, "Statistical-Physics-Based Reconstruction in Compressed Sensing," *Phys. Rev. X*, 2:021005, May 2012.
[13] S. Hamed Hassani, N. Macris, and R. Urbanke, "Thresholds Saturation in Spatially Coupled Constraint Satisfaction Problems," J. Stat. Mech. P02011, 2012.
[14] C. H. Yeung and D. Saad, "Competition for Shortest Paths on Sparse Graphs," *Phys. Rev. Lett.*, 2012.
[15] C. H. Yeung, D. Saad, and K. Y. M. Wong, "From the Physics of Interacting Polymers to Optimizing Routes on the London Underground," *Proc Natl. Academy of Sci.*, vol. 110, 2013, pp. 13717–22
[16] C. De. Bacco *et al.*, "Shortest Node-Disjoint Paths on Random Graphs," *J. Stat. Mech.*, P07009, 2014.

Biographies

David Saad is professor of information mathematics at Aston University, UK. He received a BA in physics and a BSc in electrical engineering from Technion, an MSc in physics and a Ph.D. in electrical engineering from Tel-Aviv University. He joined Edinburgh University in 1992 and Aston in 1995. His research focuses on the application of statistical physics methods to several fields, including neural networks, error-correcting codes, multi-node communication, network optimization, routing, noisy computation, and advanced inference methods.

Chi Ho Yeung received a B. Sc., M. Phil., and Ph.D. degrees in physics from the Hong Kong University of Science and Technology in 2004, 2006, and 2009, respectively. He is currently a lecturer at the Hong Kong Institute of Education. His major research interests include statistical physics, disordered systems, optimisation, routing, recommendation algorithms, complex and social networks.

Georgios Rodolakis graduated in electrical and computer engineering from Aristotle University, Greece, and obtained a D.E.A. and Ph.D. from Ecole Polytechnique and INRIA, France, in 2006. After a research fellowship at Macquarie University, Sydney, Australia, he joined CERTH, Greece, in 2011. His main research interests are in the areas of mobile networks, information theory, design and analysis of algorithms.

Dimitris Syrivelis received a bachelor degree from Technical University of Crete in 2003 and a Ph.d. from the University of Thessaly in 2009 on h/w-s/w co-designed systems for massively parallel architectures. He has developed several networking systems for wired and wireless applications, most notably for wireless network coding, information-centric networking, and software-defined networking. He is a regular contributor to open source networking projects (Click Modular Router, Linux Kernel) and he currently has 18 publications in well-known conferences.

Iordanis Koutsopoulos is assistant professor in the Department of Computer Science, Athens University of Economics and Business (AUEB), and is also affiliated with CERTH. He obtained the M.S and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park (UMCP) in 1999 and 2002, respectively. His research interests are in the broad area of network control and optimization, with applications to wireless networks, social networks, smart grid control, sensor networks, and cloud computing systems.

Leandros Tassiulas is a professor in the Department of Electrical Engineering and the Yale Institute of Network Science

at Yale University. His research interests are in the field of computer and communication networks with emphasis on fundamental mathematical models and algorithms of complex networks, architectures and protocols of wireless systems, sensor networks, novel Internet architectures, and experimental platforms for network research.

RUEDIGER L. URBANKE (EPFL) is principally interested in the analysis and design of iterative coding schemes, which allow reliable transmission close to theoretical limits at low complexities. He is a co-author of the book *Modern Coding Theory* (Cambridge University Press), a co-recipient of the 2002 and the 2013 IEEE Information Theory Society Paper Award, the 2011 IEEE Koji Kobayashi Award, as well as the 2014 IEEE Hamming Medal.

EMILIO LEONARDI is an associate professor at Politecnico di Torino. He has co-authored over 150 papers in the area of telecommunication networks. He has participated in several national and European projects, and has been involved in several consulting and research projects with private industries. He has been the coordinator of the European 7th FP-STREP "NAPA-WINE" on P2P streaming applications. His research interests are in the field of social networks, network science, and caching systems.

PAOLO GIACCONE received the Dr.Ing. and Ph.D. degrees in telecommunications engineering from the Politecnico di Torino, Italy, in 1998 and 2001, respectively. He is currently an assistant professor in the Department of Electronics and Telecommunications at Politecnico di Torino. His main area of interest is the design of network algorithms, the theory of interconnection networks, and the performance evaluation of telecommunication networks through simulative and analytical methods.

ACCEPTED FROM OPEN CALL

Rethinking the Role of Interference in Wireless Networks

Gan Zheng, Ioannis Krikidis, Christos Masouros, Stelios Timotheou, Dimitris-Alexandros Toumpakaris, and Zhiguo Ding

ABSTRACT

This article re-examines the fundamental notion of interference in wireless networks by contrasting traditional approaches to new concepts that handle interference in a creative way. Specifically, we discuss the fundamental limits of the interference channel and present the interference alignment technique and its extension of signal alignment techniques. Contrary to this traditional view, which treats interference as a detrimental phenomenon, we introduce three concepts that handle interference as a useful resource. The first concept exploits interference at the modulation level and leads to simple multiuser downlink precoding that provides significant energy savings. The second concept uses radio frequency radiation for energy harvesting and handles interference as a source of green energy. The last concept refers to a secrecy environment and uses interference as an efficient means to jam potential eavesdroppers. These three techniques bring a new vision about interference in wireless networks and motivate a plethora of potential new applications and services.

INTRODUCTION

Resources (e.g. time, frequency, code) have to be shared by multiple users in wireless networks. Therefore, interference has long been considered as a deleterious factor that limits the system capacity. In conventional communications systems, the design objective is to allow users to share a medium with minimum or no interference. Thus, great efforts are made to avoid, mitigate, and cancel interference. For instance, to support multiple users, orthogonal access methods in time, frequency, code as well as spatial domains have been used in different generations of cellular systems. In order to improve the coverage and the capacity in future-generation heterogeneous networks that will likely contain a large number of uncoordinated low-power nodes such as femtocells, interference needs to be mitigated in multiple domains, rendering its management a challenge.

Interference mitigation/avoidance techniques provide convenient mechanisms to allow multiple users to share the wireless medium. However, they lead to inefficient use of wireless resources. One may ask whether to cancel or mitigate interference is always the optimal way of utilizing wireless resources. Indeed, there has been growing interest in exploiting interference to improve the achievable rate, the reliability, and the security of wireless systems. Recently, new views on interference have resulted in advanced interference-aware techniques, which, instead of mitigating interference, explore the potential of using interference. We present two examples from the literature to illustrate the ideas.

In his early work on dirty paper coding [1], Costa proved the striking result that interference known at the transmitter but not at the receiver does not affect the capacity of the Gaussian channel. The optimal strategy to achieve this interference-free capacity is to code along interference, while canceling interference is strictly sub-optimal. Another example is coordinated multipoint or multi-cell coordination, where base stations (BSs) cooperate to serve their own and out-of-cell users.

In the downlink, the cooperating BSs work together to jointly optimize the transmitter strategies such as power, time, and beamforming design to control the inter-cell interference. Cell-edge users who suffer most from the inter-cell interference now benefit most from this coordination. In the uplink, joint decoding is performed in BSs, so signals from users in other cells are no longer treated as interference, but as useful signals.

The purpose of this article is to re-examine the notion of interference in communications networks and introduce a new paradigm that considers interference as a useful resource. We first give an overview from the information theoretic standpoint as a justification for rethinking the role of interference in wireless networks. We then introduce interference alignment and signal alignment as effective means to handle interference and increase the achievable rates. Departing from this traditional view, we present three novel techniques of interference exploitation that aim to improve the performance of wireless networks. The first technique is a data-aided precoding scheme in the multiuser downlink that judiciously makes use of the interference among users as a source of useful signal energy. In the

G. Zheng is with University of Essex. He is also affiliated with Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg

I. Krikidis and S. Timotheou are with University of Cyprus.

C. Masouros is with University College London.

D.-A. Toumpakaris is with University of Patras.

Z. Ding is with Lancaster University.

0163-6804/14/$25.00 © 2014 IEEE

second technique, we consider simultaneous information and energy transfer; in such a system, while interference links are harmful for information decoding, they are useful for energy harvesting. Thus, a favorable trade-off is demonstrated. The third technique leverages interference in physical layer secrecy as an effective way to degrade the channel of the eavesdropper and increase the system's secrecy rate.

Interference from the Information Theoretic Standpoint

We first present an overview of results on interference from information theory. The Interference Channel (IC) models simultaneous transmission by non-cooperating transmitters and receivers. The messages of each link are encoded only by the corresponding transmitter, and the receiver does not have access to the signals of other receivers. Figure 1a depicts the *K*-user Gaussian Interference Channel (G-IC). Each of *K* transmitters wants to send a message to the corresponding receiver. Receiver *k* bases its decision on signal Y_k, which contains not only the (scaled) useful signal X_k, but also interference and Gaussian noise.

Despite its apparent simplicity, to this date it is not known what the optimal way of transmitting over the G-IC is, not even for the two-user G-IC shown in Fig. 1b. Nevertheless, significant progress has been made recently, and results from information theory have started influencing the design of wireless networks. The optimal decoding strategy depends on the power of interference compared to the direct links. Interference should be treated as noise when it is very weak. The exact conditions for the two-user G-IC to be in the very weak regime can be found in [2]. In information theoretic terms, the messages of both transmitters are *private*, since they are only decoded at the intended receiver. On the other hand, when the power of the interfering signal exceeds the power of the signal of interest (*strong* interference), the optimal strategy is to also decode interference at the receivers. In this case, both messages are *public*. If the power of the interference exceeds an even higher threshold, the G-IC is in the *very strong* interference regime and the rate that can be achieved by each link is the same as if the interferer did not exist, that is, interference does not impact the achievable rates. Nevertheless, the receiver does need to decode interference in addition to the signal of interest. Clearly, there are costs associated with interference-aware decoding. The receivers are more complex, synchronization is essential, and each receiver needs to estimate not only its own channel, but also the cross-channel coefficients.

The most challenging situation arises when the power of the interference is of the order of the power of the signal of interest. To this date it is not known what the best way to transmit and decode is. A strategy that combines public and private messages (the so called Han-Kobayashi scheme [2]) achieves higher rates compared to treating interference as noise or avoiding it via orthogonal transmission, or attempting to decode all messages at each receiver. Moreover, it has been shown that as the signal-to-noise ratio (SNR) grows to infinity, a simplified Han-Kobayashi scheme can attain the capacity region within 1/2 bit [2]. In addition to providing evidence that strategies based on the Han-Kobayashi scheme may be the best for the two-user G-IC, this result may prove useful in future wireless networks with small cell size that will operate at high SNRs and will therefore be limited by interference rather than by noise.

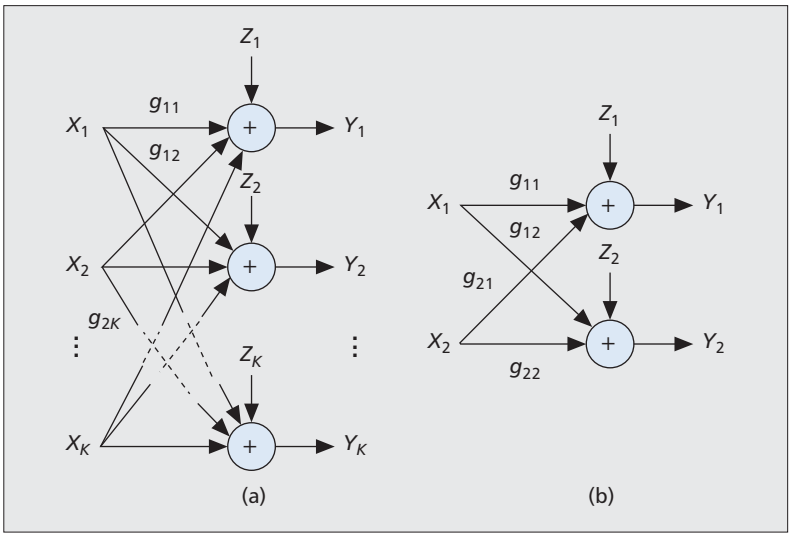


Figure 1. a) The K-user Gaussian Interference Channel (G-IC); b) the two-user G-IC.

Devising good strategies for the *K*-user G-IC seems to be even more challenging, and the Han-Kobayashi scheme does not appear to extend to the *K*-user G-IC in a straightforward manner. A promising direction toward finding good strategies for the *K*-user G-IC appears to be dealing with the combined interference by all $K - 1$ users at each receiver instead of decoding separately the interference by each user. Furthermore, a deterministic approximation framework has been developed for the G-IC, which enables the construction of structured codes [2]. By employing structured lattice codes, which are also used in other scenarios, such as multi-way relay channels, it is possible to attain the capacity region of the G-IC within a constant gap [3]. Very recently, there has been an interesting finding that connects topological interference management and index coding [5]. This connection can be leveraged to calculate rate regions that are within a constant gap from capacity and to develop transmission schemes over wireless networks. The existing index coding solutions are then translated to interference management solutions via a family of elegant achievability schemes of interference alignment (IA) that has generated significant interest, and is discussed in more detail below.

System designs that operate based on the best known achievability schemes of information theory being the ultimate goal, in the meantime improvements in performance can also be attained by incorporating interference-aware schemes in current systems. In [6] it is shown

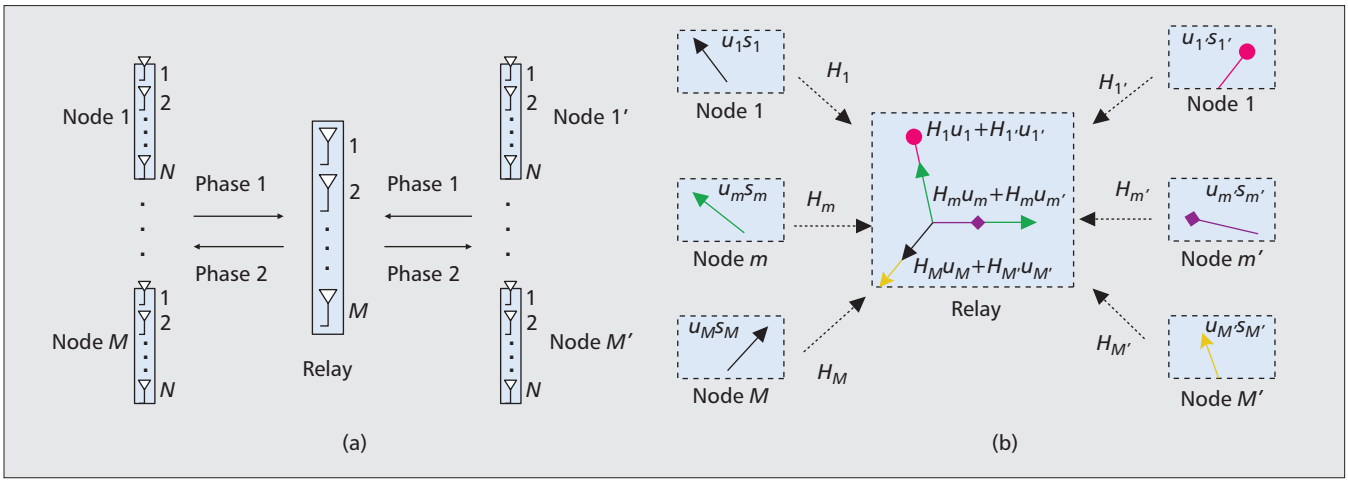


Figure 2. Illustration of the concept of signal alignment: a) system diagram for multi-way relaying, $N \le M$; b) precoding design to ensure interference alignment.

that when the transmitters use discrete constellations and interference-aware detectors are employed at the receivers, the achievable rates over the fading G-IC are limited by the SNR rather than by the signal-to-interference-plus-noise ratio (SINR).

Interference and Signal Alignment

Prior to the invention of IA [4], interference avoidance has been achieved by relying on the use of orthogonal frequency or time channels. And when interference is inevitable, conventional approaches are to adopt advanced decoding/detecting algorithms by treating interference as noise.

The success of IA lies in the fact that it efficiently exploits the rich degrees of freedom available from the time/frequency/spatial domains. By a careful coordination among the transmitters, the use of IA can ensure that all the interference is aligned together to occupy one half of the signal space at each receiver, leaving the other half available to the desired signal. As a result, the per-user rate achieved by IA for the interference channel with K pairs of single-antenna transceivers is $C(SNR) = 1/2\log(SNR) + o(\log(SNR))$. This result is surprising since a traditional view is that such a K-user scenario is interference-limited and hence the per-user rate is diminishing by increasing the number of users. As a result, the use of IA ensures that the spectral efficiency of wireless communications can be improved significantly since more users sharing the same bandwidth yields a larger system throughput.

In addition to the interference channel, the concept of IA has also been applied to other communication scenarios, including the multiple access channel, the broadcast channel, the one/two-way relaying channel, as well as physical layer security. In practice, the implementation of IA is not trivial since global channel state information at each transmitter (CSIT) is required, which is challenging, particularly for the case with fast time-varying channels. Two types of approaches to realize IA in practice have been proposed. One is to apply advanced feedback techniques, and existing results have demonstrated that the number of fedback bits needs to be proportional to the SNR in order to achieve nearly optimal performance [9]. The other is to exploit the coherent structure of channels and apply manipulations analog to space time coding at the transmitters. As a result, the concept of IA can be implemented even when the channel information is not available to the transmitters.

The concept of signal alignment can be viewed as an extension of IA in the context of bi-directional communications [10] and [11]. For example, consider a multi-pair two-way relaying communication scenario as shown in Fig. 2, where M pairs of source nodes exchange information with their partners via the relay. Each source node is equipped with N antennas, and the relay has M antennas. As can be seen from Fig. 2, the relay observes $2M$ incoming signal streams, and needs to have at least $2M$ antennas in order to separate these signals. The use of signal alignment is to effectively suppress intra-pair interference and reduce the requirement on the number of antennas at the relay. Particularly, by carefully designing the precoding vectors at the sources, the intra-pair interference is aligned at the relay, which means that the original $2M$ signal streams are merged into M streams. As a result, a relay with only M antennas can accommodate $2M$ incoming signals, which is particularly important for practical scenarios where nodes are equipped with a limited number of antennas. At the user end, each receiver can first subtract its own information, the so-called self-interference, and then detect the information from its partner, a method analogous to network coding.

Data-Aware Interference Exploitation for Multiuser Transmission

The a priori knowledge of interference is readily available at downlink transmission, where CSIT combined with the knowledge of all data symbols intended for transmission can be used to

explicitly predict the resulting interference between the symbols. Despite the insights in [1], the majority of existing precoding implementations attempt to eliminate, cancel, or pre-subtract interference. Only recently, however, has there been growing interest in making use of the interference power to enhance the useful signal [7, 8]. Indeed, it has been shown that interference can contribute constructively to the detection of the useful signal, and this phenomenon can be utilized in the CSIT-assisted downlink transmission and other known-interference scenarios to improve performance without raising the transmit power.

To clarify the above fundamental concept, a trivial example of two users is shown in Fig. 3a, where we define the desired symbol as x_1 and the interfering symbol as x_2. Without loss of generality we assume that these belong to a Binary Phase Shift Keying (BPSK) constellation and that $x_1 = 1, x_2 = -1$. For illustration purposes, we assume a lossless channel from the intended transmitter to the receiver and an interfering channel represented by the coefficient ρ. Ignoring noise, the received signal is

$$y_1 = x_1 + x_2 \cdot \rho, \quad (1)$$

where $x_2 \cdot \rho$ is the interference. Note that this model also corresponds to multi-antenna transmission with matched filtering where the correlation between the two channels is ρ. In Fig. 3b two distinct cases are shown, depicting the transmitted (×) and received (o) symbols for user 1 on the BPSK constellation. In case i) with $\rho = 0.5$ it can be seen from (1) that $y_1 = 0.5$. The destructive interference from user 2 has caused the received symbol of user 1 to move toward the decision threshold (imaginary axis). The received power of user 1 has been reduced and its detection is prone to low-power noise. In case ii), however, for $\rho = -0.5$ (1) yields $y_1 = 1.5$, and hence the interference is constructive. The power received has been augmented due to the interference from user 2 and now its detection is tolerant to higher noise power (n_{constr} compared to n_{orth}). It should be stressed that in both cases the transmit power for each user is equal to one. Note that while the above example refers to a two-user transmission scenario for illustration purposes, the fundamental concept can be extended to more users, multipath transmission, inter-cell interference, and other generic interference-limited systems.

Clearly, there are critical gains to be drawn from the exploitation of constructive interference in interference-limited transmission. As a first step, analytical constellation-dependent characterization criteria for systematically classifying interference to constructive and destructive have been derived in [7, 8] and references therein for PSK modulation. Early work carried out on a simple linear precoding technique has reported multi-fold increases in the received SNR for fixed transmit power compared to zero-forcing (ZF) beamforming [7]. This can be non-trivially translated to multi-fold savings in transmit power for a fixed received SNR. A representative result is shown in Fig. 4a where the required SNR per transmit antenna in a cellular downlink for an uncoded symbol error rate (SER) of 10^{-2} is shown for increasing numbers of single-antenna users. The results compare the widely known ZF precoding with the interference exploitation precoding of [7] for QPSK and 8PSK modulation. Significant SNR gains of up to 10dB (a 10-fold transmit power reduction) can be observed between the two techniques, by simply exploiting the existing constructive interference.

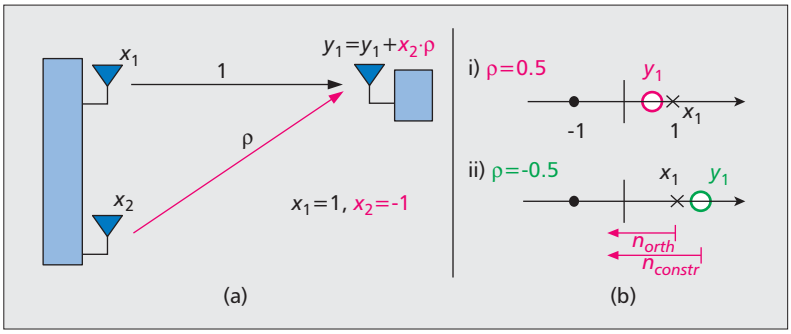


Figure 3. The concept of constructive interference — a two-user example.

Further work has investigated the application of this concept on advanced nonlinear precoding, yielding further significant gains in the transmit power. More recent work has extended this concept to inter-cell interference exploitation in multi-cellular transmission scenarios [8]. The important feature in all the above techniques is that the performance benefits are drawn not by increasing the transmit power of the useful signal, but rather by reusing interference power that already exists in the communications system, a source of green signal power that with conventional interference cancellation techniques is left unexploited.

Wireless Information and Energy Transfer

Energy harvesting (EH) communications systems that can scavenge energy from a variety of natural sources (solar, wind, etc.) for sustainable network operation have attracted significant interest. The main limitation of conventional EH sources is that they are weather-dependent and thus not always available.

A promising harvesting technology that could overcome this bottleneck is radio frequency (RF) energy transfer where the ambient RF radiation is captured by the receiver antennas and converted into a direct current voltage through appropriate circuits (rectennas). The concept of RF-EH is not new; over 100 years ago, Nicola Tesla proved and experimentally demonstrated the capability of transferring energy wirelessly. The integration of RF-EH technology into communications networks opens new challenges in the analysis and design of transmission schemes and protocols. Multi-user interference, which is the main degradation factor in conventional networks, can be viewed as useful energy signals that could be exploited for harvesting purposes. Although from an information theoretic standpoint the same signal can be used for both decoding and EH, due to practical

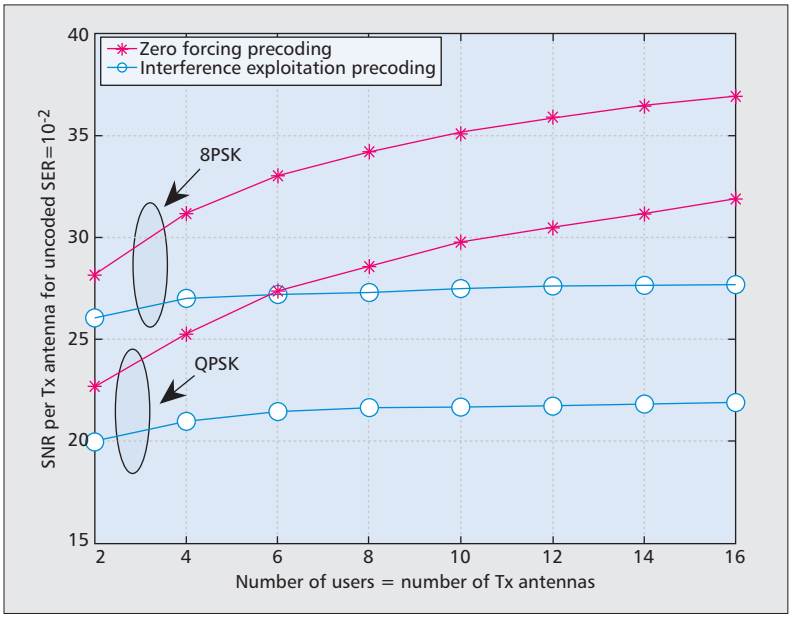


Figure 4. Required SNR per transmit antenna for an uncoded SER of 10^{-2} with increasing numbers of users and transmit antennas.

hardware constraints, simultaneous energy and information transmission is not possible with existing rectenna technology. Two practical receiver approaches for simultaneous wireless power and information transfer are "time switching" (TS), where the receiver switches between decoding information and harvesting energy; and "power splitting" (PS), where the receiver splits the received signal in two parts for decoding information and harvesting energy, respectively [12].

An interesting implication of the PS technique is that in multiuser networks, harvested energy at a particular receiver can emanate either from sources that intentionally transmit toward that direction or from other sources whose signal is traditionally perceived by that receiver as interference. Nonetheless, in this case the contribution of useful and interfering signals toward the satisfaction of any RF-EH requirements is equally important. This implication changes completely the design philosophy of such networks, as interference becomes useful.

This concept was demonstrated for the multiple-input single-output (MISO) interference channel where K transmitters, each with K antennas, communicate with K single-antenna receivers [13]; each receiver is characterized by both quality-of-service (QoS) and RF-EH constraints, while PS is used for simultaneous information/energy transfer. The QoS constraint requires the SINR to be higher than a given threshold, while the RF-EH constraint requires the power input to the RF-EH circuitry to be above a threshold. In this framework, an interesting non-convex optimization problem arises in selecting the beamforming weights and the power of the transmitters as well as the power splitting ratios at the receivers so as to minimize the total transmit power. The problem can be solved optimally using semidefinite programming, while traditional beamformers can be employed to obtain suboptimal but low-complexity solutions. An interesting conclusion is that for ZF beamforming there always exists a unique, optimal, closed-form power allocation.

The benefit of exploiting interference in the context of RF-EH is illustrated in Fig. 5, which depicts the transmit power ratio between ZF and optimal beamforming for varying SINR and RF-EH thresholds ($K = 8$). The figure indicates that by exploiting interference, the transmit power can be significantly reduced, especially for low SINR. The reason is that for low SINR there is room to increase interference, which is beneficial for RF-EH. In contrast, high SINR thresholds require almost full cancellation of interference; hence, the solutions obtained from ZF are almost optimal. The benefits of interference exploitation can also be seen with respect to the RF-EH constraints: when the RF-EH threshold increases, the ZF/optimal power ratio increases because the optimal scheme manages interference better. However, the effect of the SINR constraint on the transmit power ratio is more significant compared to the RF-EH constraint.

Interference-Aided Secrecy Rate Improvement

Due to the growing number of wireless applications, confidentiality and secret transmission has become an increasingly important issue. Recently, securing wireless communications at the physical (PHY) layer has been studied as a complimentary measure to upper layer cryptographic techniques. In the presence of eavesdroppers who passively overhear the communications, intentional interference plays a key role to improve the secrecy rate. This is understandable since interference will affect both systems; however, if properly designed, it can be an advantage for the legitimate system. This is indeed true, as shown in [14] that the exploration of aggregated interference together with location and channel quality information, can significantly improve network secrecy. In the following, we review several approaches that utilize interference to confuse the eavesdropper in a simple point-to-point network.

Consider a basic three-node system that consists of a source S, a destination D, and an eavesdropper E. When S has multiple antennas, it can transmit an information-bearing signal to D in the range space of the channel to D and also generate artificial noise (AN) to E in its null space simultaneously. In this way, even without knowledge of the instantaneous CSI of the eavesdropper, the generated AN does not interfere with the legitimate receiver D and only affects the eavesdropper node E. The same principle applies if there are trusted helper relays who could form distributed beamforming to transmit cooperative jamming signals to E.

When neither multiple antennas at S nor trusted helpers are available, the system must rely on itself to achieve secure communications. To this end, a self-protection scheme has been proposed that adopts full-duplex (FD) operation at D to improve the secrecy rate [15], as shown in Fig. 6. More specifically, an FD receiver is introduced that simultaneously receives its data while

transmitting a jamming signal to confuse E. The proposed approach uses intentional interference at D to confuse the eavesdroppers and does not require external helpers or data retransmission. Due to the FD operation, the receiver experiences a loop interference (LI) introduced by the transmitted jamming signal. If D has multiple transmit or receive antennas, it can employ joint transmit and receive beamforming for simultaneous signal detection, suppression, and intentional jamming.

In Fig. 6 the achievable secrecy rate is evaluated against the transmit SNR. We simulate two cases: single transmit/receive-antenna receiver and eavesdropper; and the receiver has two transmit and two receive antennas while the eavesdropper has four antennas for fairness. For the single-antenna case, it is seen that the FD scheme outperforms the half-duplex (HD) operation for transmit SNR greater than 10 dB, and double secrecy rate is achieved in the high SNR region. The performance of the HD scheme saturates when the transmit SNR is higher than 20 dB. When the receiver has multiple antennas and the eavesdropper adopts a simple maximal-ratio combining (MRC) receiver, the secrecy rate strictly increases with the transmit SNR and does not saturate at high SNR as in the HD case. When the eavesdropper is aware of the FD operation at D and adopts the minimum-mean-square-error (MMSE) receiver to mitigate the jamming signals from D, the achievable secrecy rate saturates at a high SNR of 40 dB but is still significantly higher than the case with HD receiver. This reveals the great potential of using interference at the receiver side to provide self-protection against eavesdropping.

Conclusions

In this article we have introduced radical views on interference in wireless networks. Traditional interference mitigation techniques are no longer optimal, and innovative ways of utilizing interference are emerging. As more aggressive resource sharing and tighter cooperation are foreseen in future wireless networks, interference management will continue to be a growing challenge.

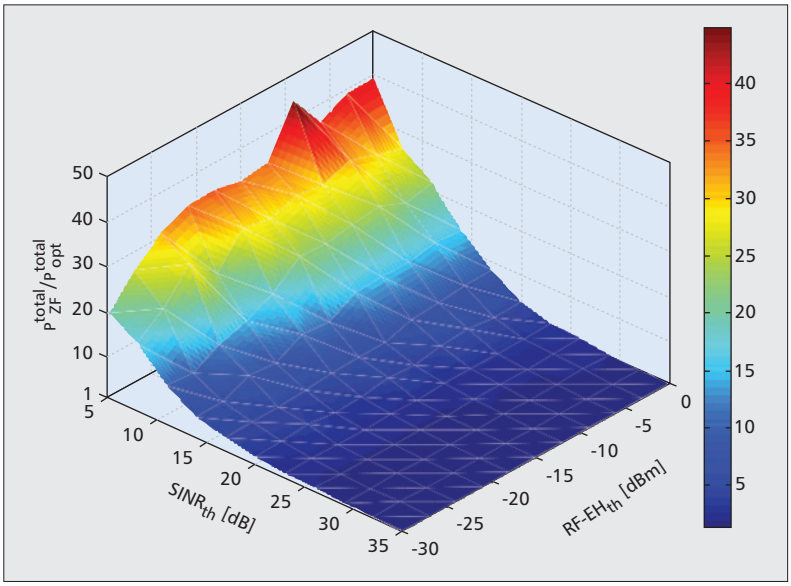


Figure 5. Transmitted power benefit from optimal exploitation of interference compared to ZF beamforming to achieve SINR and RF-EH constraints in the MISO interference channel.

Accordingly, it is essential to further these new perspectives on interference for more efficient radio resource utilization in advanced wireless concepts such as large-scale antenna arrays (massive MIMO), multicell cooperation, cognitive radio, and heterogeneous networks. Indeed, the employment of massive MIMO in future networks allows the mitigation of interference using simple linear operations. In this way, interference could be "available" in the network for other purposes without affecting its performance; this scenario motivates new services and applications. In future cloud radio access networks, baseband processing will be shifted from the BSs to the central baseband unit pool to jointly process data to and from multicells, and this offers great opportunities to fully utilize interference. In cognitive radio, the interference from the secondary user to the primary user can facilitate RF energy transfer and be tuned into useful signals if the primary data are known at

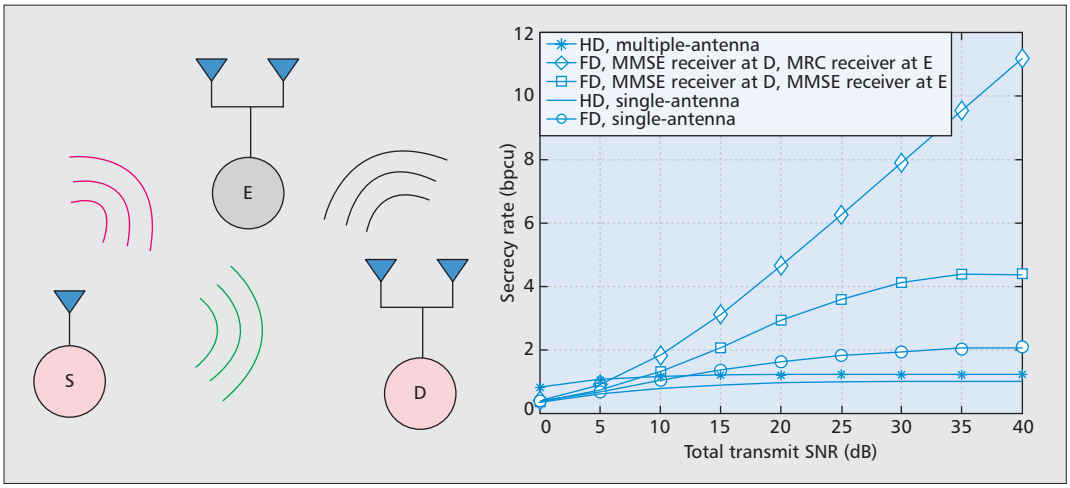


Figure 6. Left: FD operation at the receiver that creates self-interference to improve the secrecy rate; right: achievable secrecy rate in bits per channel use (bpcu) vs transmit SNR in dB.

Traditional interference mitigation techniques are no longer optimal, and innovative ways of utilizing interference are emerging. As more aggressive resource sharing and tighter cooperation are foreseen in future wireless networks, interference management will continue to be a growing challenge.

the secondary user. Regarding security in heterogeneous networks, a promising direction is to study how network interference can be engineered to best benefit wireless network secrecy.

Acknowledgment

This work was partially supported by the Research Promotion Foundation, Cyprus, under the project KOYLTOYRA/BP-NE/0613/04 "Full-Duplex Radio: Modeling, Analysis and Design (FD-RD)."

References

[1] M. Costa, "Writing on Dirty Paper," *IEEE Trans. Inf. Theory*, vol. IT-29, May 1983, pp. 439–41.
[2] A. El Gamal and Y. H. Kim, *Network Information Theory*, Cambridge University Press, 2012.
[3] B. Nazer and M. Gastpar, "Reliable Physical Layer Network Coding," *Proc. IEEE*, vol. 99, Mar. 2011, pp. 438–60.
[4] S. A. Jafar, "Interference Alignment: A New Look at Signal Dimensions in a Communication Network," *Foundations and Trends in Communications and Information Theory*, vol. 7, Issue 14, 2010, pp. 1–13.
[5] S. A. Jafar, "Topological Interference Management Through Index Coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, Jan. 2014, pp. 529–68.
[6] J. Lee, D. Toumpakaris, and W. Yu, "Interference Mitigation via Joint Detection," *IEEE JSAC*, vol. 29, no. 6, Jun. 2011, pp. 1172–84.
[7] C. Masouros, "Correlation Rotation Linear Precoding for MIMO Broadcast Channel," *IEEE Trans. Sig. Process.*, vol. 59, no. 1, pp. 252-262, Jan. 2011.
[8] C. Masouros *et al.*, "Known Interference in Wireless Communications: A Limiting factor or a Potential Source of Green Signal Power?" *IEEE Commun. Mag.*, vol. 51, no. 10, Oct. 2013, pp.162–71.
[9] R. T. Krishnamachari and M. K. Varanasi, "Interference Alignment Under Limited Feedback for MIMO Interference Channels," *IEEE Trans. Sig. Process.*, vol. 61, no. 15, Aug. 2013, pp. 3908–17.
[10] N. Lee, J.-B. Lim, and J. Chun, "Degrees of Freedom of the MIMO Y Channel: Signal Space Alignment for Network Coding," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, Jul. 2010, pp.3332–42.
[11] Z. Ding and H. V. Poor, "A General Framework of Precoding Design for Multiple Two-Way Relaying Communications," *IEEE Trans. Sig. Process.*, vol. 61, no. 6, Mar. 2013, pp.1531–35.
[12] R. Zhang and C. K. Ho, "MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, May 2013, pp. 1989–2001.
[13] S. Timotheou *et al.*, "Beamforming for MISO Interference Channels with QoS and RF Energy Transfer," *IEEE Trans. Wireless Commun.*, vol. 13. no. 5, May 2014, pp. 2646–58.
[14] A. Conti *et al.*, "Interference Engineering for Heterogeneous Wireless Networks with Secrecy," *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2013, pp. 308–12.
[15] G. Zheng *et al.*, "Improving Physical Layer Secrecy Using Full-Duplex Jamming Receivers," *IEEE Trans. Sig. Process.*, vol. 61, no. 20, Oct. 2013, pp. 4962–74.

Biographies

GAN ZHENG is currently a Lecturer at the School of Computer Science and Electronic Engineering, University of Essex, UK. He received a B. E. and an M. E. from Tianjin University, Tianjin, China, and a Ph.D. degree in electrical and electronic engineering from The University of Hong Kong, Hong Kong, in 2008. He worked as a research associate at University College London, UK, and the University of Luxembourg, Luxembourg. His research interests include cooperative communications, cognitive radio, physical-layer security, and full-duplex radio. He was the first recipient of *IEEE Signal Processing Letters* Best Paper Award in 2013.

IOANNIS KRIKIDIS received a diploma in computer engineering from the Computer Engineering and Informatics Department (CEID) of the University of Patras, Greece, in 2000, and the M.Sc. and Ph.D. degrees from Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2001 and 2005, respectively, all in electrical engineering. From 2006 to 2007 he worked as a post-doctoral researcher, with ENST, Paris, France, and from 2007 to 2010 he was a research fellow at the School of Engineering and Electronics at the University of Edinburgh, Edinburgh, UK. He is currently an assistant professor in the Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus. His current research interests include information theory, wireless communications, cooperative communications, cognitive radio, and secrecy communications.

CHRISTOS MASOUROS is currently a lecturer in the Dept. of Electrical & Electronic Eng., University College London. He received his diploma in electrical and computer engineering from the University of Patras, Greece, in 2004, an MSc. by research and a Ph.D. in electrical and electronic engineering from the University of Manchester, UK in 2006 and 2009, respectively. He has previously held a research associate position at the University of Manchester, UK, and a research fellow position in Queen's University Belfast, UK. He holds a Royal Academy of Engineering Research Fellowship 2011-2016. His research interests lie in the field of wireless communications and signal processing with particular focus on green communications, large scale antenna systems, cognitive radio, interference mitigation techniques for MIMO, and multicarrier communications.

STELIOS TIMOTHEOU holds a Dipl.-Ing from the Electrical and Computer Engineering School of the National Technical University of Athens, and an M.Sc. and Ph.D. from the Electrical and Electronic Engineering Department of Imperial College London. He is currently a research associate at the KIOS Research Center for Intelligent Systems and Networks of the University of Cyprus. His research focuses on the modeling and solution of problems that arise in communication systems, intelligent transportation systems and disaster management by developing optimization, machine learning and computational intelligence techniques.

DIMITRIS TOUMPAKARIS is an assistant professor in the Department of Electrical and Computer Engineering at the University of Patras, Greece. He holds a diploma in electrical and computer engineering from the National Technical University of Athens, Greece, and a M.S. and Ph.D. in electrical engineering from Stanford University. His current research interests include baseband system design, interference management, and multiuser information theory. He has been an editor for *IEEE Communications Letters* since 2012.

ZHIGUO DING is a chair professor at the School of Computing and Communications, Lancaster University, UK. His research interests are game theory, cooperative and energy harvesting networks, and statistical signal processing. He was co-chair of the WCNC-2013 Workshop on New Advances for Physical Layer Network Coding, and is serving as an editor for *IEEE Transaction on Communications*, *IEEE Wireless Communications Letters*, *IEEE Communication Letters*, and the *Journal of Wireless Communications and Mobile Computing*. He received the best paper award at the IET Comm. Conf. on Wireless, Mobile and Computing in 2009, was named by *IEEE Communications Letters* as an Exemplary Reviewer in 2012, and was awarded the EU Marie Curie Fellowship for 2012-2014.

CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE

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

BACKGROUND

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

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

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

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

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

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

SUBMISSION GUIDELINES

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

SCHEDULE FOR SUBMISSIONS

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

GUEST EDITORS

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

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

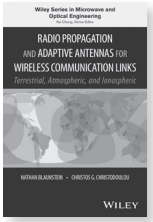
Yacine Ghamri-Doudane
Univ. of la Rochelle, France
yacine.ghamri@univ-lr.fr

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

ADVERTISERS' INDEX

Discover the Latest Titles in RF/Microwave Technology from Wiley!

Radio Propagation and Adaptive Antennas for Wireless Communication Networks, 2nd Edition

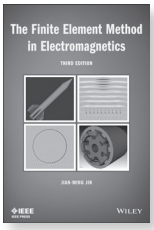


9781118659540 Cloth $175.00 5/12/2014

Nathan Blaunstein, Christos G. Christodoulou

With an emphasis on antennas and propagation, *Radio Propagation and Adaptive Antennas* investigates every aspect of wireless communication network design and function. The book delves into, among other applicable radio propagation topics, multipath phenomena, slow and fast fading, free-space propagation, and obstructed reflection and diffraction. Pertinent applications and relatable examples make this the essential modern reference for engineering practitioners and students in wireless communication systems.

The Finite Element Method in Electromagnetics, 3rd Edition

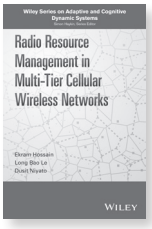


9781118571361 Cloth $175.00 3/10/2014

Jianming Jin

Useful in analyzing electromagnetic problems in a variety of engineering circumstances, the finite element method is a powerful simulation technique. This book explains the method's processes and techniques in careful, meticulous prose. It covers not only essential finite element method theory, but also its latest developments and applications. *The Finite Element Method* is an engineer's key to solving boundary-value problems.

Radio Resource Management in Multi-Tier Cellular Wireless Networks

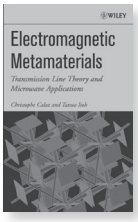


9781118502679 Cloth $115.00 12/9/2013

Ekram Hossain, Long Bao Le, Dusit Niyato

Providing an extensive overview of the radio resource management problem in femtocell networks, this invaluable book considers both code division multiple access femtocells and orthogonal frequency-division multiple access femtocells. In addition to incorporating current research on this topic, the book also covers technical challenges in femtocell deployment, provides readers with a variety of approaches to resource allocation and a comparison of their effectiveness, explains how to model various networks using Stochastic geometry and shot noise theory, and much more.

Electromagnetic Metamaterials: Transmission Line Theory and Microwave Applications

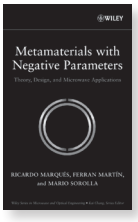


9780471669852 Cloth $150.00 8/5/2013

Christophe Caloz, Tatsuo Itoh

Electromagnetic Metamaterials: Transmission Line Theory and Microwave Applications fills an important niche, connecting the more theoretical nature of negative index materials to the practical and covers all of the important topics relevant to a very complete description of the transmission line model of negative index materials. It also includes outgrowth applications developed by the authors during their research.

Metamaterials with Negative Parameters: Theory, Design and Microwave Applications

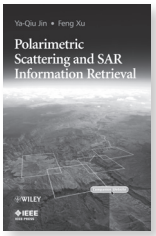


9780471745822 Cloth $120.00 8/5/2013

Ricardo Marqus, Ferran Martn, Mario Sorolla

Metamaterials with Negative Parameters approaches metamaterials using physics principles and discusses microwave applications in a uniform textbook-like manner. It provides a thorough presentation of the theory, design, and applications of metamaterials with an emphasis on split ring resonators (SRRs). The book covers all important microwave applications including filters, multiplexers, couplers, antennas, and devices.

Polarimetric Scattering and SAR Information Retrieval

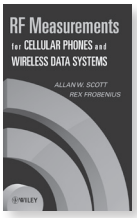


9781118188132 Cloth $140.00 7/22/2013

Ya-Qiu Jin, Feng Xu

An innovative look at Synthetic Aperture Radar (SAR), this practical reference fully covers new developments in SAR and its various methodologies, enabling readers to interpret SAR imagery. It includes theoretical scattering models and SAR data analysis techniques, and presents cutting-edge research on theoretical modeling of terrain surface. The book also covers quantitative approaches for remote sensing, such as analysis of the Mueller matrix solution of random media, mono-static and bistatic SAR image simulation. Moving clearly from fundamentals to advanced topics, this is a thorough treatment for both academic learning and independent study.

RF Measurements for Cellular Phones and Wireless Data Systems

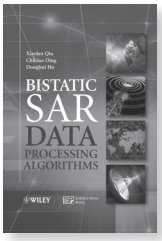


9780470129487 Cloth $147.00 7/15/2013

Allen W. Scott, Rex Frobenius

Covering all topics needed to effectively test radio frequency (RF) components and systems for cell phones and wireless data systems, this guide balances practical real-world information with relevant theory. The text summarizes basic RF principles before describing the digital technology used in cell phones and wireless data systems. Methods and equipment used in mass testing of components during manufacturing also receive detailed treatment. Industry professionals building, installing, and maintaining cell phone and wireless equipment, as well as advanced students, will soon rely on this thorough guide as a practical mainstay.

Bistatic SAR Data Processing Algorithms

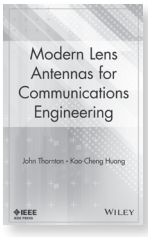


9781118188088 Cloth $150.00 6/17/2013

Xiaolan Qiu,Chibiao Ding, Donghui Hu

Focusing on imaging aspects of bistatic Synthetic Aperture Radar (SAR) signal processing, this book covers resolution analysis, echo generation methods, imaging algorithms, imaging parameter estimation, and motion compensation methods. The book is ideal for researchers and engineers in SAR signal and data processing, as well as those working in bistatic and multistatic radar imaging, and in the radar sciences. Graduate students with a background in radar who are interested in bistatic and multistatic radar will find this book a helpful reference.

Modern Lens Antennas for Communications Engineering



9781118010655 Cloth $125.00 4/1/2013

John Thornton, Kao-Cheng Huang

The aim of this book is to present the modern design principles and analysis of lens antennas. It gives graduates and RF/Microwave professionals the design insights needed to make full use of lens antennas. Because this topic has not been thoroughly publicized, its importance is underestimated. As antennas play a key role in communication systems, recent development in wireless communications would indeed benefit from the characteristics of lens antennas, namely their low profile and low cost.

1 (877) 762-2974 North America
+ 44 (0) 1243 843294 in Rest of World
Log on to www.wiley.com/IEEE

IEEE
WILEY

SAMSUNG

THE NEXT BIG THING IS HERE™

Samsung GALAXY Note® 4