January 2015, Vol. 53, No. 1

IEEE Communications MAGAZINE

www.comsoc.org

- Ambient Assisted Living Communications
- Extremely Dense Wireless Networks
- Millimeter-Wave Communications for 5G: Applications
- Ad Hoc and Sensor Networks
- Network and Service Management

IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

January 2015, Vol. 53, No. 1

IEEE Communications MAGAZINE

www.comsoc.org

- Ambient Assisted Living Communications
- Extremely Dense Wireless Networks
- Millimeter-Wave Communications for 5G: Applications
- Ad Hoc and Sensor Networks
- Network and Service Management

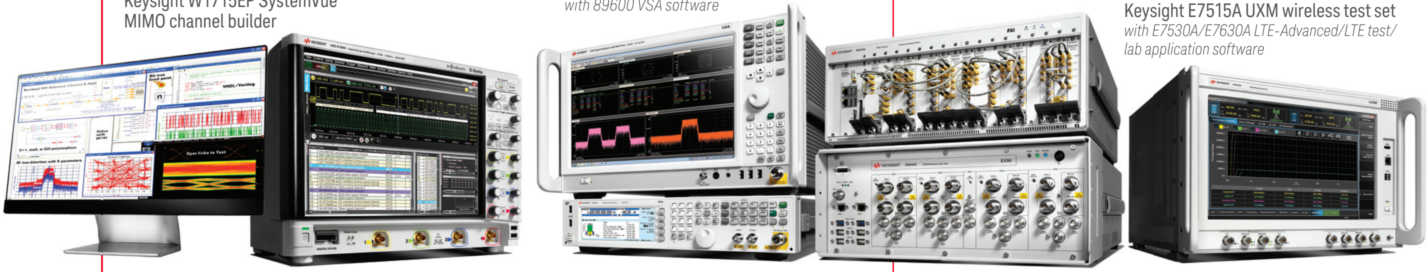
IEEE
IEEE COMMUNICATIONS SOCIETY
A Publication of the IEEE Communications Society

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

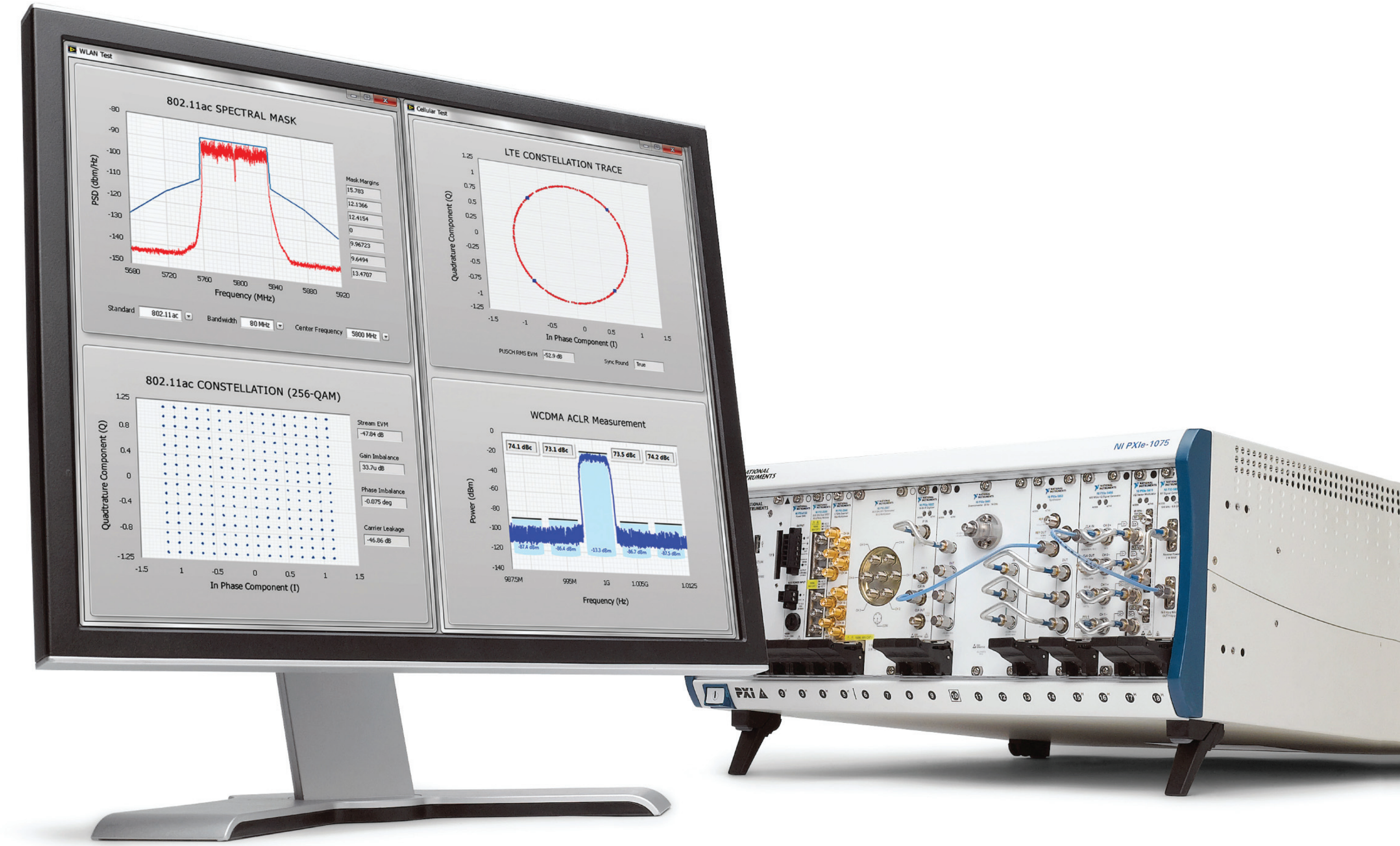
Unlocking Measurement Insights

IEEE Communications MAGAZINE

JANUARY 2015, Vol. 53, No. 1
www.comsoc.org/commag

Ambient Assisted Living Communications

Extremely Dense Wireless Networks



WLAN Test
802.11ac SPECTRAL MASK
PSD (dbm/Hz)
Frequency (MHz)
Mask Margins
Standard 802.11ac
Bandwidth 80 MHz
Center Frequency 5800 MHz
802.11ac CONSTELLATION (256-QAM)
Quadrature Component (Q)
In Phase Component (I)
Stream EVM
Gain Imbalance
Phase Imbalance
Carrier Leakage
Cellular Test
LTE CONSTELLATION TRACE
Quadrature Component (Q)
In Phase Component (I)
WCDMA ACLR Measurement
Power (dBm)
Frequency (Hz)
NI PXIe-1075

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

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

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

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

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

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

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

Millimeter-Wave Communications for 5G – Part 2: Applications

Guest Editors: Maged Elkashlan, Trung Q. Duong, and Hsiao-Hwa Chen

Ad Hoc and Sensor Networks

Series Editors: Edoardo Biagioni and Silvia Giordano

Network and Service Management

Series Editors: George Pavlou and Juergen Schoenwaelder

Seeing What Others Can't

The Key to Unlocking New Insights

Mark Wallace
Vice President and General Manager
Keysight Technologies, Inc.

You've known us as Hewlett-Packard, Agilent Technologies and, now, Keysight Technologies. For more than 75 years we have been helping you unlock measurement insights.

There have always been two sides to the story. One is the work we do, creating innovative instrumentation and software. The other is the work you do: design, develop, debug, troubleshoot, manufacture, test, install and maintain components, devices and systems.

Those seemingly separate activities are connected by something profound: the "A-ha!" that comes with a moment of insight. When those happen for us, the results are innovations that enable breakthroughs for you.

Enabling the right idea at the right time

This is our legacy. Keysight is a company built on a history of firsts, dating back to the days when Bill Hewlett and Dave Packard worked in the garage on 367 Addison Avenue in Palo Alto, California. Our firsts began with U.S. patent number 2,268,872 for a "variable-frequency oscillation generator." Appropriately, the centerpiece of Bill's design was a light bulb, which is often used to symbolize a new idea.

Our future depends on your success, and our vision is simple: by helping engineers find the right idea at the right time, we enable them to bring next-generation technologies to their customers—faster.

Offering expertise you can leverage

This is happening in aerospace and defense applications where increasingly realistic signal simulations are accelerating the development of advanced systems that protect those who go in harm's way. It's happening in research labs where our tools help turn scientific discovery into the discovery of new sciences. It's taking place with DDR memory, where our line of end-to-end solutions ranges from simulation software to protocol-analysis hardware. And in wireless communications we're providing leading-edge measurement tools and sophisticated, future-friendly software that support the development and deployment of LTE-Advanced.

Within those systems, there are more standards than a single engineer can keep up with. That's why so many of our engineers are involved in standards bodies around the world. We're helping shape those standards while creating the tools needed to meet the toughest performance goals.

Enabling your next breakthrough

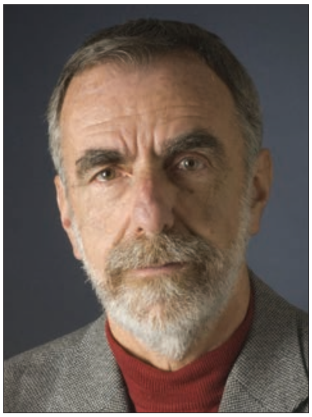
To help Keysight customers continue to open new doors, we're concentrating our effort and experience on what comes next in test and measurement. Our unique combination of hardware, software and people will help enable your next "A-ha!" moment, whether you're working on mobile devices, cloud computing, semiconductors, renewable energy, or the latest glimmer in your imagination. Keysight is here to help you see what others can't—and then make it reality.

KEYSIGHT TECHNOLOGIES

2014: A GREAT YEAR FOR COMSOC

A year has passed, and it is time for a first analysis of the main actions undertaken by ComSoc leadership. The five ComSoc Vice Presidents will describe their main achievements in 2014, but before leaving the floor to them, I would like to give a concise description of the whole picture.

2014 has been a positive year for ComSoc. After two years of heavy budgetary deficits and a 2014 forecast in the red for several hundred thousand dollars, the 2014 budget will almost break even. Moreover, the 2015 forecast is positive by approximately $100,000, which will be spent on increased services to members. The budget details are summarized in the table on the next page, showing that the 2014 forecast could be the second largest year over year improvement in the last 20 years.



SERGIO BENEDETTO

As you will see in the VPs' reports, the goals expressed in my first President's Page in January 2014 are showing good progress, and the retreat meeting in mid January 2015 will be mostly devoted to the implementation of the strategic plan that the Chair of the strategic planning committee, Byeong Gi Lee, has described in the October President's Page.

At the IEEE level, the cooperation with other presidents of large IEEE societies has been enhanced, and we have plans for the near future to collaborate in journals and conferences. In particular, we have obtained from IEEE significant advances toward a higher transparency in the IEEE and societies' budgets, and this is the first step in a path leading to a fairer revenue return to societies. As shown in the figure below, IEEE revenues have increased from $140M to $185M in five years, while the total distribution to S/Cs and MGA only grew from a bit under $60M to a bit under $70. The portion not returned to S/Cs and MGA units (going to cover costs and infrastructure) rose from approximately $80M to approximately $115M — a 43.75% in costs and an increase of 17% returned to units. This shows that there is ample margin for a fairer revenue return to societies.

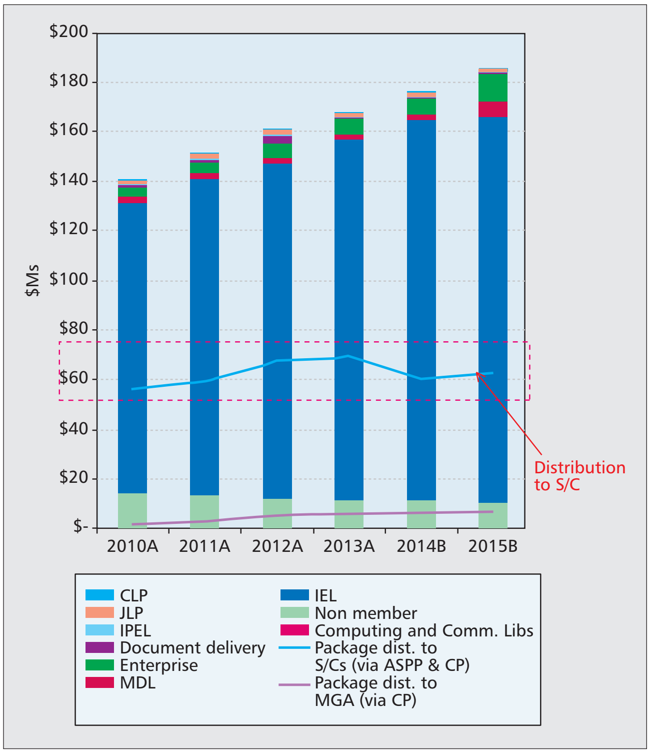


Figure 1. Select packaged product revenue and NM: trends and distributions.

Internally, ComSoc has a new Executive Director, Susan M. Brooks, replacing Jack Howell. Susan started in September 2014, and has spent her first months on the job familiarizing herself with the complex structure of a large and diversified volunteer society such as ComSoc. She has already established good and fruitful relationships with staff and volunteers, and I am sure she will prove to be a very effective ED for the society.

To conclude, ComSoc's current situation is positive, with even better prospects for the future. I wish to all ComSoc members in 2015 a year full of health, serenity, and professional success.

TECHNICAL ACTIVITIES

Technical activities is the key that ties together all activities within the Society. They not only represent the Society's cornerstone, but also its core values. In the following, we will describe some of the activities that have been conducted in 2014 as well as describe those being planned for next year.

The technical activities within the Society are managed and led by the Vice President for Technical Activities (VPTA). In this function, the VPTA is assisted by a number of senior volunteers and staff. These include Sherman Shen, Technical Activities Vice-Chair and Secretary; Michele Zorzi, Director of Education and Training; Lajos Hanzo, Chair of the Awards Committee; Zhisheng Niu, Chair of the Emerging Technologies Committee; Kin Leung, Chair of the Fellow Committee; and Steve Weinstein, Chair of the History Committee.

At the start of our term, we set out key strategic goals. These include: positioning ComSoc to target new and emerging technologies; updating and augmenting the policies and procedures vis-à-vis ComSoc technical activities; reviewing the organization and structure of our technical committees; expanding our educational offerings for professionals and practitioners, and expanding our awards program to further recognize contributions that advance the fields of interest to ComSoc.

We are pleased to report that much has been accomplished so far. Given the limited amount of space dedicated to this article, we shall briefly report on only few achievements.

2014 Accomplishments

Fellow Committee: In 2014, the committee reviewed and evaluated a total of 101 fellow nominations, the largest ever in ComSoc's history. Out of these nominees, a total of 39 members were elevated to the Grade of Fellow, the highest distinction bestowed by IEEE, only given to those engineers who have demonstrated outstanding proficiency and have achieved distinction in their profession.

Awards Committee: A number of nominations for Paper

	Budget 2015	Budget 2014	Actuals as of Nov	Forecast 2014	Actual 2013
Revenue					
Dues	1000	925	950	950	944
Member Subscriptions	355	420	348	348	378
Non Member Subscriptions	535	621	522	522	604
ASPP	2091	1897	1,483	1,897	2,439
Advertising	1100	950	825	875	959
Conferences	8910	9869	7649	9043	8787
Conference Pub Proceeds	1734	1546	797	1546	1971
Certification/Training	346	345	196	225	237
Joint Pubs	268	286	58	420	400
Page Charges/Open Access	400	325	496	530	337
All Other	79	40	26	40	23
Total	**16,817**	**17,224**	**13,350**	**16,396**	**17,079**
Expense					
Publications	1766	1430	1281	1575	1610
Advertising	363	127	204	250	248
Marketing	187	150	129	150	160
Certification/Training	235	287	152	170	292
Conferences	6997	7945	5919	7177	7187
Volunteer Travel	385	387	199	350	329
Committee Mtg Expenses	264	166	230	300	304
Programs	284	339	137	230	262
ComSoc Infrastructure	4313	4654	3929	4529	4435
IEEE Infrastructure	1882	1705	1580	1705	3310
All Other					13
Total	**16,675**	**17,190**	**13,760**	**16,436**	**18,150**
Net from Operations	**142**	**34**	**(410)**	**(40)**	**(1,071)**

Awards, Service Awards, and Career Awards were evaluated. In particular, a greater emphasis has been placed on promoting a high-integrity process while putting greater focus on transparency and openness. In addition and in order to expand our awards program to further recognize individual contributions and distinction, two new ComSoc awards have been approved by IEEE and will be awarded starting in 2015: the IEEE Communications Society Education Award, and the IEEE Communications Society Young Author Best Paper Award.

Distinguished Lecturers Program: This is a popular program that selects a distinct group of ComSoc experts to provide lecture tours in response to requests from our ComSoc chapters around the world. In 2014, 62 nominations were evaluated and a total of 16 new distinguished lecturers were selected. In particular, this year a greater emphasis was put on increasing the number of female lecturers and experts from industry.

Emerging Technologies: In order to keep abreast of the latest technology developments and evolution, volunteers are encouraged to form emerging technology subcommittees in developing technologies and areas, within the Society field of interest. This is critical if we wish to position ComSoc to target new and emerging technologies. In 2014 the emerging technologies committee had a very busy agenda and we are considering the creation of two new subcommittees, one on Big Data Processing, Analytics, and Networking, and another on Tactile Internet.

Education and Training Program: The Education and Training Board has been engaged in several activities that have been progressing very well. One major achievement is the successful completion of ComSoc's efforts to have Telecommunications Engineering as a new ABET accreditation field.

The Way Forward

In 2014, under the technical activities portfolio, many accomplishments have occurred, and this would not have happened without the hard work and dedication of so many volunteers and staff. In the year ahead we will continue our work to bring into fruition our initiatives. In the following, we will describe some of the initiatives we are planning to conduct in 2015.

Online Education and Training: One of our key goals is to develop a world-class training and professional education program that provides high quality instruction, at reasonable cost and with easy access, to address the career needs of working professionals. To do so, we are planning to extend our existing training and education programs and platforms by fostering the development of online content and training courses in communications.

Student Summer School: There is no doubt that students are the future of our Society and as such, it is of fundamental importance that we provide special membership development opportunities for them. To do so, we are planning to launch the first ever ComSoc Student Summer School in July 2015, whose aim is to: provide high-quality courses on selected topics in our field; engage local chapters in membership development and educational activities; and link our distinguished lecturers to relevant membership development activities.

Technical Committees Review: In order to continue to be at the forefront of technological development, ComSoc must evolve and continue to be at the heart of communications technology development in the world. In particular, it must reinvent itself so as to address the new challenges and opportunities. To do so, it is of fundamental importance that our technical committees play a significant leadership role. To further enhance the role of our technical committees, a review will be conducted to evaluate the technical committees' mission and structure and recommend how they can be better positioned to help ComSoc achieve its mission and grow our values to members from academia and industry.

CONFERENCES

Since the beginning of 2014 the Conferences leadership team has made a large number of accomplishments in all areas, including conference operations, conference development, conference publications, as well as in the planning and the quality of our flagship conferences. The Conflict of Interest Policy was updated, and the Steering Committee Charters, which describe the operation of the conference steering committees, as well as the composition and terms of office of their chair and members, are now in place for all of our fully sponsored portfolio conferences. Another major achievement in conference operations concerns technically co-sponsored conferences. Our society receives from 80 to 100 technical co-sponsoring (TCS) requests per year, and the process for analyzing, accepting, or declining TCS requests has been significantly improved. We now have a rigorous process and an integrated TCS document, which was published in October 2014.

On the conference development side, we have made four major accomplishments. The first is our involvement in the IEEE-IEEMA INTELEC Conference, whose 2015 edition will be held in Mumbai, India, on 22–24 January 2015. ComSoc is a financial co-sponsor of this conference together with two other IEEE Societies (CS and PES). The second accomplishment is the launch of the new IEEE Conference on Network Function Virtualization and Software Defined Networks (IEEE NFV-SDN), which is planned for 18–21 November 2015 in the San Francisco Bay Area. Other achievements include our financial participation in NetSoft 2015, which was initiated by the IEEE Future Directions Committee (FDC), and in the IEEE World Forum on Internet of Things (WF-IoT). NetSoft 2015 will be held in London on 13–17 April 2015, and WF-IoT 2015 will be held in Milan in December 2015.

A continued challenge in conference publications is the post-conference processing leading to the inclusion of conference papers in IEEE Xplore and other data bases. This process includes work by the ComSoc Staff, the volunteers (conference TPC Chair and Publication Chair), and the paper processing vendor staff. After problems experienced with GLOBECOM 2013 leading to excessive delays, a substantial improvement was achieved at ICC 2014, and a new process is now being implemented for GLOBECOM 2014 to further improve the timelines. Our goal is to not only get quality research to our conference attendees, but also to get that research into the public domain as quickly as possible after it is presented.

A new rotation plan for our flagship conferences was defined and implemented, ensuring that each region (The Americas, EMEA, and Asia-Pacific) will have an ICC or a GLOBECOM every 18 months. According to this plan, the first GLOBECOM outside the US since 2002 will be held in Singapore in 2017. Next, after ICC 2018 in Kansas City, GLOBECOM 2018 will be held in the EMEA region. The site selection process for that conference is to be completed in January 2015.

Our major objectives for 2015 include the completion of the conference portfolio analysis that started in 2014, and the establishment of a long-term vision and action plan. The GIMS and GITC will finalize their joint work on reducing the time from submission to publication, and they will define new timelines for ICC and GLOBECOM. The GITC will update the list of technical symposia and will continue to improve the quality of the paper review process. Ethical issues encountered in some recent conferences will be investigated to find solutions and avoid as much as possible this type of problem in the future.

We also plan to introduce The ICT Futures Event, which is a "new-style" event. It will commence as a series of one-day industry sector events drawing practitioners and employees from IT, networking, and communications departments in industry sectors, such as the financial sector, as well as the utilities and transportation sector. In addition, the event will introduce session formats unique to what exists in our current portfolio of conferences. As ComSoc continues to attract more industry to its conferences, the Industry Content and Exhibition Committee (ICEC) will continue to play a vital role in providing our major conferences with industry content and consultation.

In conclusion, less than 10% of ComSoc members attend a ComSoc conference annually. We will continue to ensure that all our members our aware of our conferences; there is a relevant conference for as many members as possible; and we provide as many ways as possible for our members to attend a ComSoc conference annually.

MEMBER RELATIONS

Beginning my term as VP–Member Relations, I indicated five strategic directions, envisioned as the Golden Pentagon: globalization, academia, industry, women, and students. In this first year we were able to launch and maintain several initiatives in those areas.

Chapters and Regional Activities

My priority is to favor participation of members from all regions in ComSoc activities, striving to reach out to all countries and paying extra attention to disadvantaged areas.

ComSoc regional conferences are one of the key enablers in this strategy. When they have good a reputation and attract people from both local chapters and other regions, they are a precious opportunity for members in disadvantaged areas to attend first-rate conferences, but without high expenses for traveling. Among them I wish to mention in particular the IEEE Latin-American Conference on Communications (IEEE LATINCOM). The sixth edition was recently held in Cartagena de Indias, Colombia, on November 2014, attracting authors and attendees from around the world, owing to a high quality technical program and world class keynote speakers. After six years LATINCOM has become a regular appointment for students and engineers in Latin America.

Moreover, we just organized a North-America Regional Chapter Chair Congress (NA-RCCC) at GLOBECOM 2014 in Austin, Texas, USA. I had the pleasant opportunity to meet 20 Chapter Chairs, enthusiastic and strongly motivated volunteers who spend a significant amount of time serving ComSoc. RCCCs are the chief opportunity for Chapter Chairs to meet each other, compare their experiences, and express their wishes to ComSoc leaders.

Distinguished Lecturer Program

Among ComSoc's programs for chapters and members, the Distinguished Lecturer/Speaker Program is definitely one of the most acclaimed. This year we extended the travel cost limit refundable by ComSoc from $2000 to $3000 (with the same total budget per year), thus facilitating longer tours, and ultimately looking at those areas that are seldom visited by lecturers because of their geographic location. Our motto has been "Even less DLTs, but better DLTs!"

IEEE Global Communications Newsletter

In our Member Relations strategy, the *IEEE Global Communications Newsletter* (GCN)has a key role. In order to truly make it the "Voice of the Chapters," I pursued three lines of action:

- We started to email issues, as soon as published, to all Chapter Chairs (July 2014).

- We worked to make its contents more interesting, by inviting articles and by launching a new series of monthly interviews by the VP with MR Directors (Sept. 2014).
- We totally redesigned its graphic layout, giving it a fresh and contemporary look (Oct. 2014).

Moreover, a new Online Edition of the GCN, with additional interactive contents for Chapter Chairs and links to social media, is under study and will be launched in 2015.

Student Competition

This year we conducted the second ComSoc Student Competition, "Communications Technology Changing the World," addressed to graduate and undergraduate students. I am the Co-Chair of this program with the past VP-MR Nelson Fonseca. The competition is now a well-established program in Comsoc and will be be continued in the future. In 2014 the competition was a great success. More than 70 projects were submitted and evaluated by an international committee made of 40 ComSoc experts. The evaluation process took into account social impact, technical content, originality, practical applicability, results, and quality of presentation. The winners received their Award at GLOBECOM 2014 in a Plenary Session.

Women in Communications Engineering (WICE)

The mission of this Standing Committee, chaired by Octavia Dobre, is to promote the visibility and roles of women communications engineers. Among its major initiatives in 2014 were: establishing the WICE Awards; organizing a Women's Workshop; launching a new web site and Facebook page; and finally, compiling a mailing list with more than 4000 subscribers.

Octavia and her committee are now working on a number of new initiatives. I strongly support these efforts and I cordially invite all women communications engineers to follow Octavia and stay tuned for news!

PUBLICATIONS

In the Publications area we have been active incubating new periodicals and ensuring the quality of existing periodicals. We are also expanding our online content. All this would not be possible without the Director of Journals, Len Cimini; the Director of Magazines, Steve Gorshe; and the Director of Online Content, Elena Neira.

Two new journals are launching in 2015: *IEEE Transactions on Molecular, Biological and Multiscale Communications* (TMBMC), and *IEEE Transactions on Cognitive Communications and Networking* (TCCN). TMBMC is co-sponsored by the Computer Society and the Nano Council. TCCN is co-sponsored by the Signal Processing Society and the Computer Society. We are currently incubating a journal and a magazine. In December we published the inaugural issue of the *Communications Standards Supplement* in *IEEE Communications Magazine*. This supplement was launched in cooperation with the VP of Standards. In 2015 we will launch the JSAC: Green Communications and Networking Series. We are very excited about the new publications. New publications reflect our growing field and offer opportunities and new venues of information for our readers.

This year the IEEE Communications Society's magazines and journals were reviewed by the IEEE Periodicals Review and Advisory Committee (PRAC). The PRAC reviews are an opportunity to gather information about the periodicals and receive valuable feedback from a select group of IEEE reviewers. Based on the PRAC and our own review process, we believe our journals and magazine are healthy. We want to ensure that we have quality venues for dissemination of cutting-edge information. If you have any suggestions for improving our processes, please let us know.

This year we have expanded the amount of online content to better serve our members. In addition to *IEEE Communications Technology News* and Best Readings, we have also introduced ComSoc Beats, a set of video interviews of members. If you have suggestions for interview subjects or online content, please let us know. Our aim is to provide online content that is meaningful and relevant.

STANDARDS ACTIVITIES

For 2014 the goal for my term as VP–Standards Activities was to reflect the full range of interests of our members in pre-standardization, standardization, and post-standardization activities. In addition, we stressed catching up to our field's fast moving technology initiatives in the areas of Software-Defined Networks/Network Functions Virtualization (SDB/NFV) and Internet of Things (IoT). We accomplished these with the dedicated help of Mehmet Ulema, Director of the ComSoc Standards Development Board (CSDB), and Alex Gelman, Director of the ComSoc Standards Program Development Board (CSPDB).

Rapid Reaction Standardization Process

In ComSoc, standardization occurs through the participation in research groups, study groups, and working groups, with each activity leading to the next. In 2014 we pioneered the use of a Rapid Reaction Standardization methodology, which starts from CSPDB members contacting relevant ComSoc Technical committees and asking for recommendations of experts in relevant areas. Then a face to face session is organized for one to two days and, based on the analysis of our experts, appropriate research, study, or working groups are proposed.

In April 2014 we conducted a Rapid Reaction Standardization session on SDN/NFV, and as a result created two research groups

- Software Defined and Virtualized Wireless Access
- SDN/NFV Structured Network Objects

and two study groups

- Security, Reliability, and Performance of Software Defined, and Virtualized Ecosystem
- Service Virtualization.

Similarly, in September 2014 we conducted an activity on IoT, and the decision was made to organize the following research groups

- IoT Architectures
- IoT Services
- IoT Communications and Networking Infrastructure
- IoT Security and Privacy

and a study group

- IoT APIs and Interfaces

For 2015 we expect these groups to continue. Those who are interested in any of these activities can join the Research Groups by applying for corresponding community memberships at ComSoc communities site: http://community.comsoc.org/groups.

Standards Working Groups

Our Standards Committees continued to make progress in 2014. New standards from our Dynamic Spectrum (1900.X), NGSON (1903.X), SIEPON (1904.X), Nanoscale (1906.X), and Power Line (2030.X) standards committees and working groups were in their final stages of preparation or published this year. In addition, new initiatives that have come into the CSDB include interesting work on a secure Biometric Open

STEFANO BREGNI
VP-MEMBER RELATIONS

ROB FISH
VP-STANDARDS ACTIVITIES

KHALED BEN LETAIEF
VP-TECHNICAL ACTIVITIES

HIKMET SARI
VP-CONFERENCES

SARAH KATE WILSON
VP-PUBLICATIONS

Protocol (P2410.X) and on high speed, uncompressed Audio/Video networking, known as HDBaseT (P1911.X).

Standards Publications

One of the signal achievements in the Standards Activities area in 2014 (with the cooperation and support of our VP–Publications, Katie Wilson) was the organization and publication of the first *Communications Standards Supplement* in *IEEE Communications Magazine*. This supplement is intended to serve the interests of the members of the global communications and networking standards community. The first issue contains six standards-related articles and news from global Standards Development Organizations. A call for Feature Topic proposals has been issued for upcoming issues in 2015-2016. Looking toward the future, we hope that this supplement can develop into a full fledged Communications Standards Magazine.

Standards Meetings and Conferences

This year the third successful edition of our workshop, "From Research to Standards," was held at ICC 2014. Based on this success, a new ComSoc portfolio conference has just been approved: the IEEE International Conference on Standards for Communications and Networking (IEEE-CSCN). The inaugural venue of this conference series is planned for Yokohama, Japan in November 2015. Come and be a part of it!

VICE PRESIDENT BIOGRAPHIES

Stefano Bregni is an Associate Professor of Telecommunications at Politecnico di Milano, Italy. Born in 1965, he is graduated in Electronics Engineering. He worked in industry with SIRTI before joining Politecnico in 1999.

He is an IEEE Senior Member (1999). He has been an IEEE ComSoc Distinguished Lecturer for seven years. From 2003 to 2009 he visited repeatedly 14 countries and 29 IEEE Sections/Chapters worldwide, giving more than 30 lectures in Mexico, Puerto Rico, El Salvador, Panama, Costa Rica, Guatemala, Ecuador, Peru, Bolivia, Colombia, Malaysia, India, Poland, and USA.

In ComSoc he has served as: Vice-President Member Relations (2014–15); Board-of-Governors Member-at-Large (two terms: 2010–12, 2013); Director Education (2008–11); Chair of the Transmission, Access and Optical Systems TC (2008–09; Secretary/ViceChair 2002–07); and Member-at-Large of the GLOBECOM/ICC Technical Content (GITC) Committee (2006–09).

He is ICC2016 Technical Program (TP) Co-Chair. He has been GLOBECOM2012 TP Vice-Chair, LATINCOM2011 TP Co-Chair, GLOBECOM2009 Vice-Chair for Symposia, and Symposium Co-Chair in nine other ICC/GLOBECOMs. He is the Editor-in-Chief of the *IEEE Global Communications Newsletter* and an Associate Editor of *IEEE Communications Surveys and Tutorials*. He contributed to ETSI/ITU-T standards on network synchronization. He is the author of 80+ refereed papers and of the fundamental book *Synchronization of Digital Telecommunications Networks* (J. Wiley, 2002).

Robert S. Fish [SM] received his Ph.D. from Stanford University. Currently, Dr. Fish is President of NETovations, LLC. From 2007 to 2010 he was Chief Product Officer and Senior VP at Mformation, Inc. From 1997 to 2007 he was Vice President and Managing Director of Panasonic's US R&D Laboratories. Prior to this he was Executive Director, Multimedia Communications Research, at Bellcore after starting his career at Bell Laboratories.

He has more than 30 publications and 17 patents. During his career he and his organizations have initiated and managed standards development activities in IEEE, 3GPP, OMA, IETF, ATSC, CableLabs, OSGi, SDRF, and probably a few more. He is Vice President for Standards Activities of the IEEE Communications Society. He co-edited a series in *IEEE Communications Magazine* on IEEE standards in communications and networking. He has served on the Com-Soc BoG and was Chair of GIMS. He is a member of the Board of Governors of the IEEE Standards Association, Chair of IEEE-SA's Global Coordination Committee, and a founding member of the IEEE-SA Corporate Advisory Group. He is Co-Founder and Steering Committee Chair of ComSoc's CCNC Conference. For his leadership and contributions to the Multimedia Communications Technical Committee, he was the recipient of MMTC's Distinguished Service Award.

Khaled Ben Letaief received the BS degree with distinction, MS, and Ph.D. degrees in electrical engineering from Purdue University. He is currently Dean of Engineering and Chair Professor at The Hong Kong University of Science & Technology, with expertise in wireless communications and networks. In these areas, he has published more than 470 journal and conference papers and given invited keynote talks as well as courses all over the world. He has 13 patents, including 11 US patents.

He served as a consultant for different organizations and is the founding Editor-in-Chief of the *IEEE Transactions on Wireless Communications*. He has served on the editorial board of other prestigious journals, including the *IEEE Journal on Selected Areas in Communications — Wireless Series* (as Editor-in-Chief). He has been involved in organizing a number of major international conferences, including: WCNC'07 in Hong Kong; ICC'08 in Beijing; ICC'10 in Cape Town; TTM'11 in Hong Kong; and ICCC'12 in Beijing.

He is a Fellow of IEEE and a Fellow of HKIE. He is also the recipient of many other distinguished awards, including the 2007 IEEE Communications Society Publications Exemplary Award; 2009 IEEE Marconi Prize Award in Wireless Communications; 2010 Purdue University Outstanding Electrical and Computer Engineer Award; 2011 IEEE Communications Society Harold Sobol Award; 2011 IEEE Wireless Communications Technical Committee Recognition Award; and 10 IEEE Best Paper Awards. He is recognized by Thomson Reuters as an ISI Highly Cited Researcher.

He served as an elected member of the IEEE Communications Society (ComSoc) Board of Governors, IEEE Distinguished Lecturer, IEEE ComSoc Treasurer, and IEEE ComSoc Vice–President for Conferences. He is currently serving as IEEE ComSoc Vice–President for Technical Activities, member of the IEEE Product Services and Publications Board, and member of the IEEE Fellow Committee.

Hikmet Sari is currently a Professor and Head of the Telecommunications Department at SUPELEC, near Paris, and also Chief Scientist of Sequans Communications. Prior to moving to these positions, he held various research and management positions at Philips, SAT (SAGEM Group), Alcatel, Pacific Broadband Communications, and Juniper Networks. He has served as an Editor of the *IEEE Transactions on Communications*, Guest Editor of the *European Transactions on Telecommunications*, Guest Editor of *IEEE JSAC*, Associate Editor of the *IEEE Communications Letters*, Chair of the Communication Theory Symposium of ICC 2002, Technical Program Chair of ICC 2004, Vice General Chair of ICC 2006, General Chair of PIMRC 2010, General Chair of WCNC 2012, Chair of the GITC Committee in 2010–2011, Distinguished Lecturer of the IEEE Communications Society (2001–2006), Member of the IEEE Fellow Evaluation Committee (2002–2007), and Member of the Awards Committee (2005–2007). His distinctions include the IEEE Fellow Grade and the Andre Blondel Medal in 1995, the Edwin H. Armstrong Award in 2003, the Harold Sobol Award in 2012, and election to the European Academy and to the Science Academy of Turkey in 2012.

Sarah Kate Wilson received her A.B. from Bryn Mawr College with honors in mathematics in 1979, and her Ph.D. from Stanford University in electrical engineering in 1994. She has worked in both industry and academia and has been a Visiting Professor at Lulea University of Technology, the Royal Institute of Technology in Stockholm, and Stanford University. She is currently an Associate Professor of electrical engineering at Santa Clara University. She has served as an Associate Editor for *IEEE Transactions on Wireless Communications*, *IEEE Communications Letters*, and *IEEE Transactions on Communications*, and as Editor-in-Chief of *IEEE Communications Letters*. Her research interests include orthogonal frequency division multiplexing (OFDM) and optical wireless communications. She is a Fellow of the IEEE and a former Director of Journals for the IEEE Communications Society.

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)

CONFERENCE REPORT

IEEE GLOBECOM 2014 HOSTS 57TH ANNUAL INTERNATIONAL CONFERENCE IN THRIVING ENTREPRENEURIAL AND TECHNOLOGICAL CENTER KNOWN AS "THE SILICON HILLS"

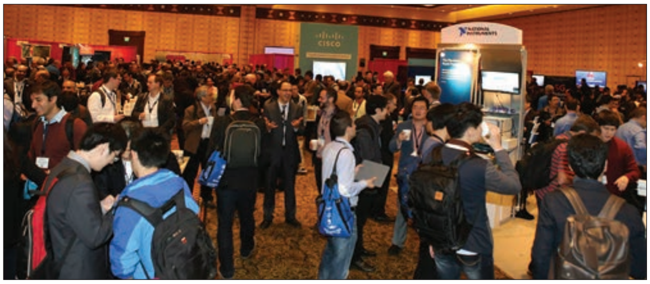
MAYOR LEE LEFFINGWELL PROCLAIMS DECEMBER 8–12 AS IEEE WEEK IN AUSTIN, TEXAS TO HONOR PREMIER EVENT DEDICATED TO GLOBAL TELECOM ADVANCEMENTS AND INNOVATIONS

IEEE
2014
GLOBECOM
GLOBAL COMMUNICATIONS CONFERENCE EXHIBITION & INDUSTRY FORUM
AUSTIN, TX USA
★ 8-12 DECEMBER 2014 ★
THE GREAT STATE OF COMMUNICATIONS

IEEE GLOBECOM 2014 (www.ieee-globecom.org/2014), the premier global conference dedicated to driving advancement in nearly every telecommunications field, held its 57th annual event from December 8–12 in Austin, Texas, the thriving national center for entrepreneurialism and innovation commonly recognized as "the Silicon Hills." Known for its high concentration of technology startups and corporations, the conference's organizing committee combined the talents of thousands of industry professionals to ensure an enriching experience for both practitioners and academia. This included expanding the industrial program with a far wider breadth of keynotes, executive forums, demos and interactive sessions as well as offering amenities such as a downloadable mobile application, which placed session and conference proceeding information, locations and times at the fingertips of attendees. AT&T and the Austin Hilton further supported IEEE GLOBECOM 2014 by assuring access to the latest 4G LTE cellular technology and state-of-the-art WiFi connections were available throughout the conference venue.

With this backdrop, Austin provided the ideal setting for this year's activities attended by nearly 3,000 industry experts, scientists and academics participating in more than 2, 500 presentations detailing the latest breakthroughs in cutting-edge fields like SDN, NFV, Small Cell Networks, HetNets, Internet of Things (IoT), Cloud Computing, Millimeter Wave MIMO, Vehicular Networks and 5G communications.

Themed "The Great State of Communications," IEEE GLOBECOM 2014 opened Monday, December 8 with the first of two full days of tutorials and workshops exploring topics like "Green Broadband Access: Energy Efficient Wireless," "Evolution Toward 5G Cellular Networks," "Emerging Technologies for 5G Wireless Cellular Networks (Wi5G, Formerly B4G)," "Mobile Communications in Higher Frequency Bands (MCHFB)" and "Optical Wireless Communications (OWC)." For instance, the day began with Dr. Matthias Illing of the Center of Competence Connectivity, Robert Bosch GmbH, Germany, offering his presentation titled "Connecting the Dots: How the IoTS Changes Multi-Domain Enterprises" during the "International Workshop on the Internet of Things and Services." As highlighted by Dr. Illing, "We are at the beginning of the commercialization of the Internet," where "75 percent of the world's population will be connected via the Internet by 2015" and "tens of billions of devices connected five to 10 years from now." This will occur in conjunction with a string of megatrends resulting in new business models and robust IT-based services that will one-day include automated driving, dynamic vehicular mapping and sensors that will automatically alert authorities to accidents and summon ambulances. However, this will require the integration of many different standardization efforts and the "creation of tools and fusion software that overcome security issues and tie everything together."


The crowded exhibit hall.

Throughout the day, other workshops explored the "Management of Emerging Networks and Services," which detailed the emergence of innovations such as sensors embedded in soap dispensers within hospital operating rooms to ensure everyone involved is properly scrubbed before procedures and the "Cloud Computing Systems and Network Applications" necessary to launch the first commercial 5G deployments in 2020 achieving superior user experiences and peak data rates with a "revolutionary clean slate design" that stores" everything on the cloud."

In the evening, IEEE GLOBECOM held its annual "Welcome Reception" in the Grand Ballroom of the Austin Hilton with hundreds of attendees enjoying live music from a six-piece honkytonk band, while dining on Texas barbeque and gourmet cupcakes. Other highlights included the opening of the conference's exhibition hall promoting the displays and interactive demonstrations of AT&T, the conference's first diamond sponsor, and other leading corporations like National Instruments, Huawei, Cisco, Samsung, Nokia, Qualcomm, Alcatel-Lucent, Ericsson, Intel, BEECube, TVC, Freescale, Fujitsu, NYU WIRELESS, Elsevier, Wiley, Springer and Cambridge University Press. Another first was the IEEE GLOBECOM 2014 Poster Session featuring the event's top 50 paper selections and topics such as "Traffic Modeling for Machine-to-Machine (M2M) Last Mile Wireless Access Networks" and "Traffic Management for Sustainable LTE Networks."

On Tuesday morning, IEEE GLOBECOM 2014 officially commenced with the remarks of Dr. Ted Rappaport; General Chair John Donovan, Senior Executive Vice President of Technology and Operations at AT&T; and Mike Rollins, President of the Greater Austin Chamber of Commerce, who welcomed everyone to this growing, technical center that is "driven by the outstanding university-based research of 40 separate institutions" and the "work of 13 percent of the population, which are employed by the city's high-tech community." In addition, IEEE President Roberto de Marca recognized the "unwavering commitment of the organization" to solve problems and improve the quality of life and planet through communications," while IEEE ComSoc President Sergio Benedetto lauded the tremendous high-quality work of the committees, staff and volunteers in making this year's event "more fulfilling to industry."

Following these comments, Dr. Edward G. Amoroso, Chief Security Officer, AT&T Inc., offered the conference's first keynote focused on "Recent Advances in Cloud Security" and an outdated approach to security that he equated to stacking sandbags against an Internet perimeter filled with "ports and protocols built into the edge and letting activity in." As a result, Dr. Amoroso proposed the building of "virtual object security architectures" composed of "micro domain rings" that would more effectively protect a "broken perimeter that doesn't worry about the gateway and servers." This would include a robust botnet-like structure offering far more resiliency, redundancy and security.

James Truchard, President, CEO and Cofounder of National Instruments, then highlighted his company's "never before seen approach to prototyping next generation wireless systems" during his keynote address on "Next-Generation Tools for Next-Generation Wireless Research." During his presentation, Dr. Truchard highlighted National Instruments' "drive to make life easier" by "accelerating the speed of innovation and ability to try new things." For example, Dr. Truchard spoke of new platforms that synchronize hardware, software and algorithm research at the system design level enabling faster implementations and "the migration to the latest technologies with great efficiency."

As a result, the LabView Communications System Design Suite was demonstrated at IEEE GLOBECOM 2014 as a revolutionary method for rapidly prototyping real-time software defined radio and wireless systems. In addition to showcasing several highly-innovative applications designed with LabView at the conference's exhibition hall, National Instruments also hosted a complimentary, half-day tutorial offering hands-on experience in designing, simulating, and deploying to hardware a prototype of a 20 MHz LTE-based real-time OFDM link on a high-performance FPGA. This was followed by transmitting data over the air using the link created by the participant to illustrate the integrated algorithm-to-FPGA flow made possible by LabVIEW Communications software and NI SDR hardware.

Afterward, IEEE GLOBECOM 2014 initiated its three-day schedule of learning sessions consisting of nearly 900 technical paper presentations, hundreds of addresses from industrial leaders, senior-executive panel discussions and numerous business and industrial forums focused on the innovations and research representing virtually every area of broadband, wireless, multimedia, data, image and voice communications. This included 10 industry panels on Tuesday alone dealing with the latest research into areas like "Emerging Technologies for Next Generation WiFi," "Programmable Carrier Infrastructures" and "CloudRAN Architectures, Vitualization and Connectivity Solutions for 5G Cellular Communications" as well as the executive forum on "Network Transformation" that described a world of "ubiquitous communications connecting everything that can be connected by 2020." According to the participants, this will be highlighted by a "new global network" with "geographic boundaries blurring everywhere," "ecosystems of computers working together" and an intelligent abstraction network of SDN/NFV solution architectures ensuring amazingly "fast, super real-time and reliable connections" that "users can modify for themselves with rapidly evolving new services."

Other highlights from the day included the Annual Awards Luncheon hosted by IEEE ComSoc Awards Chair Lajos Hanzo and IEEE ComSoc President Sergio Benedetto. This began by announcing Anthony Soong, Chief Scientist for Wireless Research and Standards at Huawei Technologies, and Wei Yu, Professor of the Electrical and Computer Engineering Department at the University of Toronto as the organization's 2014 IEEE Fellows. Additional honorees were:

- Harpreet S. Dhillon, Radha Krishna Ganti, Francois Baccelli and Jeffrey G. Andrews, who were awarded the Leonard G. Abraham Prize for their article titled "Modeling and Analysis of K-Tier Downlink Heterogeneous Cellular Networks."
- Vahid Tarokh, Hamid Jafarkhani and Robert Calderbank, who received the Advances in Communication Award for their article on "Space-Time Block Coding for Wireless Communications: Performance Results."

Best Paper Award winners.

- Kien T. Truong and Robert W. Heath, who were provided the Journal of Communications and Networks (JCN) Best Paper Award for their article titled "Effects of Channel Aging in Massive MIMO Systems."
- Andrea Goldsmith, a Professor of Electrical Engineering at Stanford University, who was provided the Edwin Howard Armstrong Achievement Award "for sustained and fundamental contributions to wireless communications."
- Wen Tong, Head of Huawei Wireless Research and Executive Vice President of Huawei Canada Research Center, who received the Industrial Innovation Award "for leadership in and contributions to 3G and 4G wireless communications systems."
- Eros Spadotto, Executive Vice President, Technology Strategy and Operations of TELUS, who was given the Distinguished Industry Leader Award "for contributions and leadership in the development of the mobile communications industry and innovation in mobile communications technology."
- Stefano Bregni, Associate Professor of Telecommunications at Politecnico di Milano, who was provided awarded the Harold Sobol Award for Exemplary Service to Meetings and Conferences "for more than 15 years of sustained outstanding personal commitment and contribution to the organization and technical management of IEEE Communication Society's flagship and regional conferences."
- Chengshan Xiao, Professor of Electrical and Computer Engineering at Missouri University of Science and Technology, who received the Joseph LoCicero Award for Exemplary Service to Publications "for extraordinary service and leadership as the Editor-in-Chief of IEEE Transactions on Wireless Communications."
- Lin-shan Lee, Professor of Electrical Engineering and Computer Science of National Taiwan University, who was given the ComSoc/KICS Exemplary Global Service Award "for contributions to ComSoc in the international activities, the development of global collaboration with sister societies, and the promotion of global volunteer participation and services."
- John M. Cioffi, Hitachi America Professor of Engineering (Emeritus) at Stanford University, was provided the 2014 IEEE Leon K. Kirchmayer Graduate Teaching Award "for educating a stellar array of graduate students in digital communications and for inspiring them to make a difference."
- Dipankar Raychaudhuri and Narayan B. Mandayam, who were given the 2014 IEEE Donald G. Fink Award for their paper "Frontiers of Wireless and Mobil Communications."
- Chapter of the Year was awarded to the Malaysia Chapter for the Asia/Pacific Region.
- Chapter Achievement Award Recipients were Republic of Macedonia for the EMEA Region, Colombia for the Latin America Region, New Jersey Coast for the North America Region and Malaysia Chapter for the Asia/Pacific Region.

The opening day of the conference concluded Tuesday night with the Dialogue with Industry Executives, which provided the

Keynote session attendees.

viewpoints of Tareq Bustami, VP of Product Management, Freescale; Rod Naphan, CTO and Senior VP, Fujitsu Network Communications; Farooq Khan, President, Samsung Research America; Bob Gessel, Head of Technology Strategy Development, Ericsson North America; and Mike Murphy, CTO, Networks, North America, Nokia. Moderated by David Lu, Vice President, Business Solutions Development, AT&T, these panelists were tasked with providing insights to wide ranging audience questions that focused on the increased role of women in science, vehicle-to-vehicle communications and the fundamental role of technology in furthering the quality of life.

Throughout the exchange, the participants accentuated the need of industry to motivate young individuals through the ongoing support of programs like STEM, internships, participation in academic advisory boards and most importantly helping them to "find the spark" in this "exciting age of innovation and human empowerment" where "a lot of good comes from chasing the lofty goal." While foreseeing roads filled with self-driving vehicles, all agreed that this future was still far off given the amount of challenges and depth of logistical, real-time and pervasive innovations needed to "guarantee nothing goes wrong." However, everyone rallied behind the pursuits of such goals as a means to "revolutionize other industries" through the creation of technologies that will help solve "real-human problems" ranging from energy, water and food shortages to terrorism.

Wednesday began with the keynotes of Pankaj Patel, Executive Vice President and Chief Development Officer at Cisco, who posed the question "Are You Ready for the Internet of Everything?" and proceeded to detail a world where 39 percent of the population is already connected to 13 billion devices and more data was generated in the past year alone than the past 5,000 years combined. Already powering 80 percent of all businesses, Patel highlighted his presentation by describing "the phenomenal impact that big digital analytics are having on what we do everyday" through examples of smart trappings that minimize the use of insecticides, smart pill bottles, manufacturing robots and buoys with sensors collecting biological information at the Great Barrier Reef in Australia. As cited by Patel, this will all be part of the "3rd wave of IT transformations" and "consumer-centric enterprises" that will "create data at a rate never seen before" and machine-driven analytics that "will not only recognize patterns, but predict anomalies."

Dr. Wen Tong followed with a discussion of "5G Wireless Beyond Smartphones" and the dramatically improved end user experiences and new revenue streams expected from the next generation of enhanced efficiencies and innovations. According to Dr. Tong, this will include shifting from the mobile Internet to the connected world by 2020 to provide six billion smartphone users 100x more connectivity, latency levels below 1ms and operational speeds of 10gbps. Driven by 55 trillion sensor readings every hour, 5G will also be punctuated by a truly orthogonally and synchronous air interface concept enabling the auto drive of 400 million cars in the next decade.

The remainder of Wednesday entailed a wide selection of technical sessions exploring topics like wireless sensor networks, video streaming, mobile cloud networking and next generation network design as well as industry panels dedicated to programmable and big data cloud networking, IPv6 and IoT challenges, and cable industry access technologies. Another highlight was the "5G Vision: Requirements and Key Technologies" executive forum that explored the "tech enablers" creating a platform in which "affordable, sustainable and ultra-reliable" wireless applications as well as "services that are not yet even known" will be made possible through the flexible, full-duplex of the licensed, unlicensed and shared licensed spectrums.

Held at the Bob Bullock Texas History Museum, the Wednesday evening banquet was accentuated by an IEEE GLOBECOM attendee-only tour of the three-story museum dedicated to Texas history and many of its noted exhibits like the 300-year-old shipwrecked remains of the LaBelle French sailing vessel and "Fly Girls," a special tribute to women aviators in WWII. After an exceptionally delicious meal, Ted Rappaport and Sergio Benedetto presented the crystal globe to representatives from IEEE ICC 2015 (London), while Dr. Ed Tiedemann, the general chair of IEEE GLOBECOM 2015 (San Diego), provided an enticing introduction to next fall's conference.

Thursday morning then began with the keynote of Dr. Alicia Abella, Assistant Vice President (AVP), AT&T Labs, who addressed "Cloud Computing: A New Strategic Infrastructure" and the formation of an industry devoted to cloud computing and the standardization of cloud technology components. She began by comparing the cloud to other disruptive undertakings such as the development of the U.S. interstate highway system and the need to overcome the fears of those frightened by the loss of control of privately-managed infrastructures. This includes fostering a "love of the cloud" that clearly delineates its vast abilities for accelerating new products to market, reducing maintenance and equipment purchase costs and "gaining efficiencies through services used only when needed." As a result, Dr. Abella foresees the continual shift from big IT departments toward cloud providers offering more network functions with increased QoS over the next few years.

Furthering the exploration of "Future of Wireless" trends was Rajesh Pankaj, Senior Vice President, Engineering at Qualcomm Research, who detailed how new 5G services and devices will connect new industries, while empowering new user experiences. "Already more prevalent than electricity or running water in some regions," 2020 will witness the shipment of eight billion smartphones enabled by 1000x higher efficiencies, 1000x more small cells, seamless coverage, flawless mobility and reliable users experiences. "All these things will then transform communities, enterprises and industry" with fiber-like, mobile broadband capabilities that deliver mission critical services like remote surgery and uniform, simultaneous connectivity.

The morning's opening keynote session ended with the annual presentation of the IEEE GLOBECOM 2014 Best Paper Awards. Chosen for this year's honors were:

- Arash Gholami Davoodi and Syed Ali Jafar, University of California Irvine for their paper titled "Settling Conjectures on the Collapse of Degrees of Freedom Under Finite Precision CSIT."
- Alexandra Bousia and Elli Kartsakli, UPC, Spain; Angelos Antonopoulos, Telecommunications Technological Centre of Cat-

alonia, Spain; Luis Alonso, Universidad Politecnica de Catalunya-Barcelona TECH and Telecommunications and Aerospatial Engineering School of Castelldefels, Spain; and Christos Verikoukis, Telecommunications Technological Centre of Catalonia, Spain for their entry on "Small Cells for Energy Efficient Networking: How much does it cost?"

•Miao Wang, Ran Zhang and Sherman Shen, University of Waterloo, Canada; Muhammad Ismail and Khalid A. Qaraqe, Texas A&M University at Qatar; and Erchin Serpedin, Texas A&M University, USA for the paper titled "A Semi-distributed V2V Fast Charging Strategy Based on Price Control."

•Lin Ye and Hongli Zhang, Harbin Institute of Technology, China for their research on "Modeling Leechers Attack in Bit-Torrent."

•Wei Zhao, Zubair Fadlullah, Hiroki Nishiyama and Nei Kato from Tohoku University, Japan and Kiyoshi Hamaguchi, NICT, Japan for their entry "On Joint Optimal Placement of Access Points and Partially Overlapping Channel Assignment for Wireless Networks."

•Hisham ElShaer, King's College London-Vodafone Group, United Kingdom; Federico Boccardi, Vodafone Group, United Kingdom; Mischa Dohler, King's College London, United Kingdom; and Ralf Irmer, Vodafone Group, United Kingdom for their entry on "Downlink and Uplink Decoupling: a Disruptive Architectural Design for 5G Networks."

•Yun Liao, Tianyu Wang, Lingyang Song, Peking University, China; and Zhu Han, University of Houston, USA for their research titled "Listen-and-Talk: Full-duplex Cognitive Radio Networks."

•Nan Cheng, Ning Lu, Ning Zhang, Sherman Shen and Jon Mark, University of Waterloo, Canada for the paper titled "Opportunistic WiFi Offloading in Vehicular Environment: A Queueing Analysis."

•Xinru Zheng, Hua Zhang and Wei Xu, Southeast University, China; and Xiaohu You, National Mobile Communication Research Lab, Southeast University, China for the submission on "Semi-orthogonal Pilot Design for Massive MIMO Systems Using Successive Interference Cancellation.

•Bo Yu and Liuqing Yang, Colorado State University, USA; and Hiroyuki Ishii, DOCOMO Innovations, Inc, Japan for the paper titled "3D Beamforming for Capacity Improvement in Macrocell-Assisted Small Cell Architecture."

•Yongxiong Ren, Long Li, Guodong Li, Yan Yan, Yinwen Cao, Hao Huang, Nisar Ahmed, Zhe Zhao, Giuseppe Caire, Andreas Molisch and Alan Willner, University of Southern California, USA; Marton Lavery and Miles Padgett, University of Glasgow, United Kingdom; Chongfu Zhang, University of Electronic Science and Technology of China, China; and Moshe Tur, Tel Aviv University, Israel for the entry titled "Experimental Demonstration of 16 Gbit/s millimeter-wave Communications using MIMO Processing of 2 OAM Modes on Each of Two Transmitter/Receiver Antenna Apertures."

•Tao Li and Pingyi Fan, Tsinghua University, China; and Khaled B. Letaief, The Hong Kong University of Science and Technology, Hong Kong for the paper "Data Acquisition with RF-based Energy Harvesting Sensor: From Information Theory to Green System."

•Jingqing Wang and Xi Zhang, Texas A&M University, USA for their submission on "3D Percolation Theory-Based Exposure-Path Prevention for Optimal Power-Coverage Tradeoff in Clustered Wireless Camera Sensor Networks."

•Song Noh, Michael Zoltowski and David Love, Purdue University, USA for their research on "Downlink training codebook design and hybrid precoding in FDD massive MIMO Systems."

These award presentations were then followed by the rapid-fire, panel discussion of representatives from Freescale, Dell, IBM Research Labs, National Instruments, Silicon Labs and Cisco offering their perspectives on the future of interconnectivity during the morning's "Internet of Things – from Standardization to Deployment and Commercialization" executive forum. According to the experts, many predict that 2020 will be earmarked by 50 billion devices, including 1.9 billion smartphones, talking to each other and sharing data through a standardized language they all understand. However, all agreed that these advances must be preceded by the development of energy harvesting chips and hardware supported by "sustainable and super long battery life," devices operating with "simple and reliable functionality every moment of the day" and adaptable architectures meshing together all layers of sensors and embedded controls to further enable the collection, analysis and proactive response to all forms of big data.

During the Thursday lunch break, conference attendees and hopeful entrepreneurs were treated to the insights of Stanford Professor Andrea Goldsmith, patent/IP expert Ryan McCarthy from Fish & Richardson and Brian Magierski from Powershift, who shared their professional and personal expertise on the challenges and rewards of "launching your own technology company." As lessons learned, Dr. Goldsmith listed several fundamental ingredients for creating startups. They included "the passion for turning an idea into a technology and then a company" and the necessity of working with co-founders, partners and investors, who have the talent and experience needed to build successful businesses. Among the biggest mistakes cited by the panelists are the failure "to understand the legal consequences of decisions at an early stage to avoid future pitfalls" and/or form advisory boards that can help increase the access to financial resources and manage the multitude of marketing, design and manufacturing risks.

On Friday, December 12, IEEE GLOBECOM 2014 concluded its comprehensive agenda with a second day of workshops and tutorials detailing specific topics such as "Wireless Networking and Control for Unmanned Autonomous Vehicles: Architectures, Protocols and Applications," "Green Radio Transmission," "Advanced RF Communications Technology for Public Safety/Homeland Security," "Optical Wireless Communications," "Vehicular Networking," "Communication-Aware Multi-Robot Control and Behavior Optimization," "Low Power Wide Area Machine-to-Machine Communications Using LTE," "Wireless Small Cell Networks: Past, Present, Future," and "Emerging Concepts and Technologies Toward 5G Wireless Networks."

For more information on IEEE GLOBECOM 2014 including access to the conference's comprehensive program and Facebook, LinkedIn and Twitter links, please visit

www.ieee-globecom.org/2014

Planning is also already underway for IEEE GLOBECOM 2015 to be held December 6 – 10 in San Diego, California. Please visit www.ieee-globecom.org/2015 for conference updates, registration details and presentation guidelines.

The IEEE GLOBECOM 2015 Call for Papers deadline is April 1, 2015 for original symposia submissions in the areas of Ad Hoc and Sensor Networks; Cognitive Wireless Networks; Communication and Information Systems Security; Communication QoS, Reliability and Modeling; Communications Software, Services and Multimedia Applications; Communication Theory; Next-Generation Networking; Optical Networks and Systems; Signal Processing for Communications; and Wireless Communications and Networking. Selected areas in communications will also cover Access Networks and Systems; Big Data Networking; E-Health; Cloud Computing; Data Storage; Green Communications and Computing; Internet of Things; Molecular, Biological and Multi-scale Communications; P2P Networking; Powerline Communications; Satellite and Space Communications; Smart Grid Communications; Social Networks and SDN & NFV.

Conference Preview

IEEE HPSR 2015: The 16th IEEE International Conference on High Performance Switching and Routing

July 1–4, 2015, Budapest, Hungary

Tibor Cinkler, HPSR 2015 General Chair, Budapest University of Technology and Ecomonics, Hungary

IEEE HPSR 2015
16th International Conference on High Performance Switching and Routing
Budapest, Hungary
July 1-4, 2015

The dynamicity of modern communication networks relays on the switching and routing infrastructure.The Cloud, the Internet of Things, the Big Data, the mobile data communications, the 4k video all demand an even more dynamic network provided by high performance switching and routing infrastructure.

The next annual IEEE HPSR conference will be highlighted by the presentations of leading communications experts addressing the latest worldwide advancements in topics of software defined networks (SDN), network function virtualization (NFV) and clouds; the convergence of fixed and mobile access with excessive offloading from mobile to Wi-Fi networks; handling the various traffic, quality, fairness and availability requirements; deploying large data centers and enhancing their switching capabilities; provisioning content delivery networks; achieving energy efficiency of switching and routing equipment. These are only a few of the topics that have demanded switching and routing capabilities that are more intelligent, more efficient, and more reliable than ever before.

The IEEE HPSR 2015 conference will host technical and business presentations highlighting the entire range of switching and routing technologies, including

- Architectures of high-performance switches and routers
- High-speed packet processors
- Address lookup algorithms
- Packet classification, scheduling, and dropping
- Switching, bridging, and routing protocols
- Latency and buffer control
- Multicasting and anycasting
- Low latency and reliable routing for industrial Internet
- Security issues in industrial Internet
- Internet of Things (IoT)
- Big Data driven networks
- P2P routing
- Routing in wireless, mobile and sensor networks
- Optical switching and routing (including spectrum elastic networks)
- Switching, bridging, and routing in data centers and clouds
- Software defined networking/network function virualization (SDN/NFV)
- Open-source routing, Open Network Control Architecture
- Data placement and migration
- Multiprocessor networks
- Network control and management
- Pricing, accounting, and charging
- QoS/QoE and scalability of switching, bridging, and routing
- Traffic characterization and engineering
- Power-aware switching, bridging, and routing protocols
- Interface selection and traffic routing in 5G FMC offloading
- Routing and switching for 5G core networks
- Protection switching and routing, restoration, availability

HPSR 2015 will include tutorials on latest advances in switching and routing in both the core and the converged fixed-mobile fronthaul/backhaul part of the 5G mobile network. As major trends the following themes are highlighted: spectrum elastic optical core networks; software defined architectures with virtualized network functions; mobile to Wi-Fi offload; the impact of IoT and BigData onto Switching and Routing infrastructure. Tutorial proposals are expected at http://www.ieee-hpsr.org.

For the first time a special Demonstration Session will be organised at the HPSR where new switching architectures, new routing and resource control approaches and softwares, and other HPSR relevant concepts can be demonstrated. The Demonstration Presenters will have an Exhibition Booth for the demonstration.The authors of demos are invited to submit their proposals along with their demo requirements. For more information, visit http://www.ieee-hpsr.org.

HPSR 1015 will include numerous Invited talks as well as presentations from industry players — vendors and operators as well as research institutions—to present their results and share their views on latest developments.

HPSR 2015 will also host panels discoursing latest advances on various aspects of switching and routing along with the impact of new Internet paradigms onto switching and routing.

For more information on IEEE HPSR 2015 including speaking, demonstraction and tutorial opportunities, registration information, and conference updates, please visit

http://www.ieee-hpsr.org

GCN
GLOBAL
COMMUNICATIONS
NEWSLETTER

January 2015
ISSN 2374-1082

MEMBER RELATIONS

Latin America Region

Interview with Pedro Aguilera, Director of the Latin America Region

By Stefano Bregni, Vice-President for Member Relations, and Pedro Aguilera, Director of the Latin America Region

This is the fifth article in the series of eight, begun in September and published monthly in the *Global Communications Newsletter*, which covers all areas of IEEE ComSoc Member Relations. In this series of articles I introduce the seven Member Relations Directors (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 they present their activities and plans.

In this issue I interview Pedro Aguilera, Director of the Latin America Region. Pedro is an Electrical Engineer, graduated from the University of Chile. He is an IEEE Senior Member and has been an active volunteer for ComSoc and IEEE since 2003. Last year he was the General Chair of the IEEE Latin American Conference on Communication (LATINCOM 2013). Currently he is a member of the IEEE Chile Section Board, Chair of the ComSoc Chile Chapter, and Director of the ComSoc Latin America Region. He has worked for 17 years on network planning and technology development at Telefónica Chile. Presently he is Account Manager at Switch Comunicaciones, Chile.

Stefano Bregni

Pedro Aguilera

It is my pleasure to interview Pedro and to have this opportunity to present the organization and activities of the Latin America Region, which I have visited dozens of times, beginning in 2003 as an IEEE Distinguished Lecturer.

Stefano: Hola Pedro! Como estás? You have been Director of the Latin America Region for three years already. What are the biggest challenges and best opportunities in your Region?

Pedro: Hola Stefano! Yes, it is true. How quickly have these three years passed! About your question: I think we can do great things in this region. We have a lot of potential, but we need to achieve a higher level of participation and integration between ourselves. We have to take advantage of the many things we have in common: our cultural roots, our language, etc. Therefore, some of my personal objectives as Director of the LA Region are to achieve greater use of telecommunication technologies; increase the use of the DL/DS Programs; improve and extend the contributions to the GCN; improve the transition process from the Regional Director whose term is expiring to the new incoming Director; making the LA Board more efficient by having more meetings and by extending its composition.

Stefano: How would you extend the use of telecommunication technologies in the ComSoc activities of the LA Region?

Pedro: I deeply believe that ComSoc should be a model for the rest of the IEEE in this matter. Webinars, for example, are of great benefit for the members of ComSoc of Latin America who live in remote areas of the big cities. It is a good way to share the numerous technical activities in the region carrying out each chapter. Just imagine if we were able to transmit online and record each technical conference, we could construct a valuable database for our current members and attract new ones.

There has been some progress in this area already. Each year two or three webinars are realized in LA. Some of them are combined with the Distinguished Lecturer Tours program. Last year we also incorporated technical conferences in Spanish and Portuguese, the two predominant languages in the region. I encourage the use of webinars for all Latin American chapters. Today we have Webex accounts available at no cost for all chapters of ComSoc. It is easy to use and the technical requirements are minimal.

Stefano: You mentioned the Distinguished Lecturer/Speaker Programs. Over my 10 DLTs in the past, six were in Latin America. So I am very much interested to know more about what are your objectives and achievements in this area.

Pedro: The Distinguished Lecturer Tour (DLT) and the Distinguished Speaker Program (DSP) are much appreciated sources of activity for members of ComSoc. However, Latin America has some particular challenges that it must be addressed. This year we tried to organize seven DLTs, but only four were finally completed. ComSoc's Distinguished Lecturers are distributed mainly in North America, Europe, and Asia. The geography of Latin America makes it very difficult to get DLs from Europe and Asia. The high cost of the airfare and the long hours of travel are difficult barriers to overcome. In this sense, in 2014 we made valuable progress on flexibility for the limits of funds for international tickets associated with Distinguished Lecturer Tours. This is very positive and we thank the MPD Director, Dr. Koichi Asatani, and of course you, Stefano, for the strong support, without which we would not have achieved this flexibility.

Stefano: Thanks a lot. It is always good to hear when someone appreciates our work. It is true indeed that I always presented the Distinguished Lecturer Program as one of the most important programs of ComSoc, because it is an effective way to serve members, especially in disadvantaged areas. Another element that I believe is central in ComSoc's Member Relations strategies is the *Global Communications Newsletter*. I am working to improve its contents and distribution. Indeed, I want it to be considered the "Voice of the Chapters." Would you tell us something about your plans about GCN in the LA Region?

Pedro: I must admit that our region contributes very little to this excellent newsletter. The number of articles sent does not reflect the level of activity that we carry on in the Region. In conjunction with our GCN Regional Correspondent, Lisandro Zambenedetti, we have been implementing some steps to improve

(Continued on Newsletter page 5)

Distinguished Lecture Tour of Ying-Dar Lin in Australia, June 2014

By Ying-Dar Lin, National Chiao Tung University, Taiwan

This was my first DLT (Distinguished Lecture Tour), piggybacked on the IEEE International Conference on Communications (ICC) in Sydney, Australia, 10-14 June 2014. The reason to piggyback this DLT with a conference trip is to save money and time. During ICC I attended several editorial board meetings, TPC meetings, and technical meetings. I also gave the first lecture at the University of Sydney, hosted by Prof. Albert Zomaya, who is the Editor-in-Chief of *IEEE Transactions on Computers*. I've known Prof. Zomaya for three years, since I started serving on his editorial board. We decided that we might have a common research interest in an emerging area (software defined networking (SDN)) that combines the IEEE Computer Society and the IEEE Communication Society.

After ICC I flew to Melbourne to give two other lectures at Deakin University, hosted by Prof. Shui Yu, and Swinbourne University of Technology, hosted by Prof. Grenville Armitage. I met Prof. Yu at a conference in Hawaii, and he invited me to give a talk at their one-day workshop on Emerging Topics in Computer Science. With Prof. Armitage, I recently finished guest-editing a successful special issue on Open Source for Networking in *IEEE Network Magazine*. This special issue attracted a record high 70 submissions, and we had to split it into two issues, published in March 2014 and September 2014.

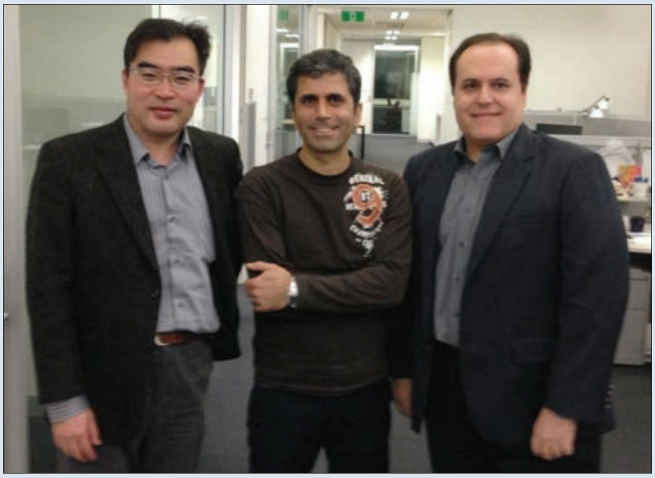
Although I provided five topics (Research Roadmap Driven by Network Benchmarking Lab, Traffic Forensics, Benchmarking Smartphones, Open Source for Networking, Software Defined Networking) for my hosts to choose from, they all picked the same topic — Software Defined Networking: Why, Where, When, and How — because SDN is an emerging area and could fundamentally change the networking industry. A four-minute introductory video to the five topics is available at https://www.youtube.com/watch?v=BuxQ9Yk3OXc&feature=youtu.be.

Special thanks should go to the chairs of two local chapters who helped arrange the local publicity for the lectures: Prof. Jinhong Yuan, New South Wales ComSoc Chapter Chair, and Dr. Paul G Fitzpatrick, Victorian ComSoc Chapter Chair.

Three Lectures

The lecture itself is a tutorial and survey on SDN. I argued why, where, and when for SDN. Then I illustrated how SDN works in four sections: standardization, development, testing, and deployment. These four sections reflect the viewpoints from standard bodies, vendors, test labs, and operators. It was final exam week in all universities in Australia. The attendees were mostly faculty members, post-doc researchers, and Ph.D. students. The number of attendees at the University of Sydney, Deakin University, and Swinbourne University of Technology was approximately 25, 50, and 30, respectively.

Left to right: Ying-Dar Lin, Leith Campbell, and Grenville Armitage after the talk at Swinbourne University of Technology.



Left to right: Ying-Dar Lin, one of Prof. Zomaya's students, and Prof. Albert Zomaya after the talk at the University of Sydney.

Attendees after the talk at Deakin University.

In-Depth Discussions after Lectures

The lectures triggered many questions from the audience. I briefly summarize their major questions and my answers below.

Why active networking failed and why SDN could succeed? Though both promote the idea of network programmability, active networking tried to put the control of that programmability into every router, i.e. running programs on routers, which is infeasible. In SDN the control of programmability is the cloud at the controllers, i.e. running programs on controllers to program routers and switches. Cost reduction and new service revenue would be the two driving forces for the success of SDN.

Why redirecting data-plane packets to controllers could lead to performance problems? Most data-plane processing at switches is done in hardware, e.g. table lookup in ASIC, but control-plane processing at controllers and applications is done in software. Redirecting data-plane packets to controllers would trigger control-plane software processing, which slows down the forwarding process. Thus, the redirection ratio should be reduced.

How can one controller serve a large network? Currently there are approximately 100,000 domains on the Internet. Some of them would be turned SDN-enabled. Each domain can have one controller or multiple controllers for fault tolerance and load balancing.

How about the Internet backbone? The evolution starts from data centers, then service providers and their subscriber networks they support, i.e. enterprise, residential, and cellular users. It may evolve into handheld and wearable devices, but the entire Internet backbone itself is likely to remain the same, i.e. running BGP routing in a distributed way, because the Internet backbone does not belong to a single domain.

How is the routing information collected in SDN with just

(Continued on Newsletter page 6)

CONFERENCE REPORT

Highlights from the Mobile Cloud Networking Workshop 2014, Lisbon, Portugal

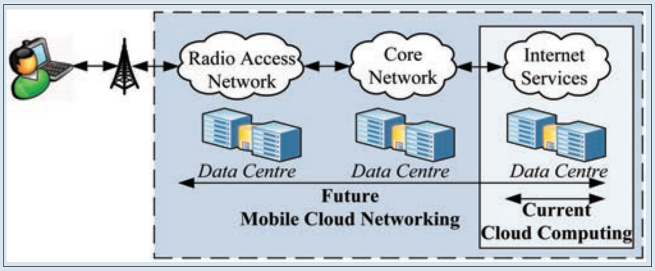
L. M. Correia and L. S. Ferreira (INOV-INESC/IST – University of Lisbon, Portugal), Paulo Simões (ONE Source, Portugal), Jorge Carapinha (PT Inovação e Sistemas, Portugal), and T. M. Bohnert (ZHAW, Switzerland)

On 19 June, 2014 the Mobile Cloud Networking workshop was held at IST – University of Lisbon, Portugal. It was co-organised by the IEEE ComSoc Portugal Chapter and MCN (Mobile Cloud Networking), an EC FP7 R&D project. It was motivated by an ongoing transformation that drives the convergence between the Mobile Communications and Cloud Computing industries, enabled by the Internet (see figure upper right). The expression of interest in this workshop was reflected by the number of participants: more than 90 professionals from industry, operators, regulators, and universities. It was a full-day workshop, with talks from 13 experts, some involved in the MCN project, but others having been invited, coming from manufacturers, operators, and universities in Portugal.

The Mobile Cloud Networking Vision

Mobile communication networks are constantly growing and being enhanced. Driven by an increase in capacity needs from end-users, these networks need to continuously densify their deployments, and upgrade and optimize their networks with the most recent systems advances. Current network elements are based on hardware, designed by vendors for specific purposes. Their deployment and configuration requires the intervention by technicians on each device, while most of the time their upgrade and re-size requires their replacement. The concept of cloud, based on data-centers with servers networked all together, makes it possible to install software that performs similar network functionalities. This enables the easy configuring and updating of network elements, and elastically scaling the associated computation, storage, and networking resources. Given a network with a small load, the software can run only on one machine, while if the load drastically increases, the system is able to increase dynamically the number of machines to support the necessary features. The concept of mobile cloud networking will change the paradigm of network operators and equipment manufacturers. From the moment operators have platforms where they can easily install software with various types of features, the use of resources becomes more dynamic, adaptive, and efficient, thereby reducing the costs of acquisition and operation. New business models will appear, giving rise to new players with new value-added services. For the end user, this certainly means more choice among operators, more services, and lower costs.

An overview of MCN, a project that aims to offer a service combining mobile networking, computing, and storage elastically, on-demand, and only paid per use, was presented by Thomas M. Bohnert and Andy Edmonds (ZHAW, Switzerland). A novel business player, the mobile cloud provider, is foreseen, driving the requirements for a mobile network architecture that exploits and supports cloud computing, and enabling the exploitation of the concept of an "end-to-end Mobile Cloud" for novel applications. A vision of the use of cloud technologies in mobile core networks was given by Rui Gomes (Vodafone, Portugal). Cloud will ease trials, full network swaps, and deployments. It is a solution based on low-cost and reusable hardware, software being the differentiator. This is good for software telco vendors and newcomers, challenging hardware-based vendor models, seen as a clear win-win for operators and vendors. New ICT consumption models in corporate environments were addressed by José Pereira (Novabase, Portugal). In a very aggressive market, where pricing models tend to be decided based on business (OPEX consumption), cloud can provide a benefit in deploying scalable and pay-as-you-grow solutions.



An ongoing transformation is driving the convergence between the Mobile Communications and Cloud Computing industries, enabled by the Internet.

MCN Workshop panelists.

The trend toward a cloudified service function chaining infrastructure was highlighted by Rui Aguiar (Instituto de Telecomunicações, Portugal). Two burning telco topics were presented: Software Defined Networking (SDN) and Network Function Virtualisation (NFV), which point toward cloud concepts. It was reinforced by Raúl Caldeira (Ericsson, Portugal), who discussed how NFV, SDN, and cloud are transforming networks, services, and business, highlighting how "it takes three to tango." NFV virtualizes network functions, Cloud scales them to the cloud for optimal deployment, and SDN enables cross-domain control, orchestration, and management, enabling infrastructure to meet changing demands. In the same direction, the talk from Fernando Carvalho (Portugal Telecom, Portugal) traveled from SDN to the Software Defined Cloud (SDC) concept, which provides management and automation of computing, networking, and storage resources through simple portals. Key requirements for the next-generation cloud are fast time-to-market, scalability, quality of service, and total cost of ownership. Emulation techniques to speed up the convergence between mobile communications and cloud computing were presented by Manuel Ricardo (University of Porto, Portugal). ns-3 may contribute to the reduction of development times of new solutions (distributed systems, protocols, resource management techniques), enabling performance evaluation and main use case validation, combination of models with real building blocks, fast prototyping, and short validation times.

From the MCN project, several innovations were presented. An architecture to offer cloud-based Radio Access Network (RAN) as a Service (RANaaS) was presented by Dominique Pichon (Orange, France), bringing cost and efficiency benefits from the cloud computing model. It aims at offering elastic, scalable, and on-demand RANaaS, dynamically adapted to geographic and temporal load variations. Several challenges were discussed concerning the front-haul, baseband unit, radio resource management, real-time performance, and scalability. On the other side, an architecture to offer IP Multimedia Subsystem (IMS) as a service for NFV-based architectures was presented by Giuseppe Carella (Univeresity of Berlin, Germany). It offers the possibility to deploy on-demand an instance of the IMS platform. With a very limited number of entities, a large number of subscribers can be served. Finally, a live demonstration of LTE Evolved Packet Core

(Continued on Newsletter page 6)

REGIONAL REPORT

Professional Development Opportunity through IEEE ComSoc Workshops in New Zealand

By Nurul I Sarkar, IEEE Joint NZ North and South ComSoc Chair

In New Zealand (NZ) we have two ComSoc Chapters. A joint chapter of IEEE NZ North and South Sections, and the other a joint Chapter of the Communications, Signal Processing, and Information Theory (COM/SP/IT) Societies of the New Zealand Central Section. We are in the process of forming a single joint ComSoc Chapter NZ-wide spanning all three sections. We believe that members of ComSoc and the wider community would benefit from this single joint Chapter.

Being a ComSoc chapter chair, Associate Professor Nurul Sarkar had organized a day-long workshop for professional development of the members of the society and the wider university community. The workshop was held on 4 November 2013 from 9 am to 4 pm at Auckland University of Technology (AUT), Auckland, New Zealand. The event was sponsored by IEEE NZ North Section and the School of Computer and Mathematical Sciences, AUT. The workshop began with a short introduction by Chapter Chair Dr. Nurul Sarkar highlighting the role of the ComSoc Chapter and its association with IEEE NZ North Section. He then introduced distinguished invited guest speakers, local presenters, and the overall program for the day.

The workshop consisted of a series of presentations on aspects of computer networking and communications. There were 12 presenters, including three invited guest speakers from the National University of Defence Technology (NUDT), China; the remaining nine presenters were from AUT's Network and Security Research Group (both staff and research students). A brief description of each presentation is highlighted below.

Shuaib Memon (Ph.D. research student) gave a talk on "Strict QoS Guarantee for Emergency Traffic in Wireless Networks" by summarizing his Ph.D. research at AUT. Akbar Hossain, another Ph.D. student, spoke about "Spectrum Management in Cognitive Radio Wireless LANs." The talk highlighted the importance of developing new techniques/algorithms for spectral efficiency. Next, Priyanka Undugodage (AUT research assistant) gave a talk on "Achieving Transmission Fairness in Wireless Mesh Networks" by highlighting opportunities and research challenges.

Among the three guest speakers from NUDT China, Professor Jibo Wei gave an interesting presentation on aspects of "Broadband Communication and Network Research" to share his research ideas. Associate Professor Haitao Zhao gave a talk on "Available Bandwidth Estimation and Prediction in Cognitive Wireless Networks" by summarizing his research activities and findings. Professor Jun Zhang talked about computational photography, which generated a lot of interest among industry participants for further discussion.

Our next presenter was Dr. Sayan Ray (Manukau Institute of Technology, Auckland) who gave an interesting talk on "Base Station Congestion Control Mechanisms in Natural Disaster Situations." The talk highlighted practical system implications. Sumeet Thakur, another research assistant, gave a talk on "Simulation and Modelling of LTE-A Using OPNET Modeler."

Among the three presenters from the academic staff, Dr. Jairo Gutierrez gave a talk on "Wireless and Mobile Network Pricing Models" by highlighting the link between technology and businesses in the global wireless and mobile markets. Krassie Petrova, another staff member, spoke about interesting aspects of a mobile learning research framework and future directions. Mee Loong (Bobby) Yang highlighted his research on "The Multiple-Key Blom's Key Agreement Scheme." Finally, Sotharith Tauch (Ph.D. research student) gave a talk on "Cascading Failure" by summarizing his work.

The length of each presentation ranged from 15 to 30 minutes. Academic staff members had slightly more time than research students to share their research and development work with the audience. Each presentation concluded with open research areas that generated a lot of interest among the participants for further discussion. The tutorial style of presentation helped the audience attain a thorough understand of the emerging research topics. There was also ample opportunity for questions and answers after each presentation.

Despite of the busy time of the year, approximately 50 people (30 IEEE members and 20 non-members) from the wider university community attended the event. There was ample opportunity for informal discussion, networking, and international collaboration, especially during morning tea and lunch break. People enjoyed the facilities provided by AUT University. There was also a discussion session in the afternoon especially for international research collaboration.

Overall it was a productive opportunity for the attendees to network, encourage academia-industry links, collaborate, and share ideas. Organizing Chair Associate Professor Nurul Sarkar received positive feedback from the participants, indicating that the event was very successful.

Informal discussion and networking during tea break.

Networking and international collaboration opportunity.

IEEE ComSoc workshop attendees in Auckland.

REGIONAL REPORT

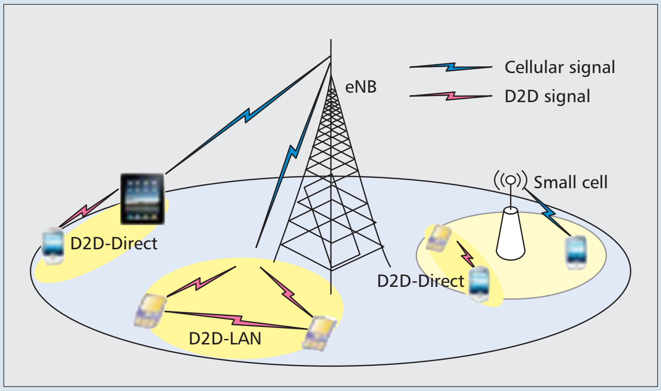
Coordinated Device-to-Device Local Area Networks: The D2D-LAN Project in China

By Lingyang Song, Peking University, Beijing, China

As the research on the fifth generation mobile communications steps into its startup period, the Ministry of Science and Technology (MOST) of China granted a five-year (2013-2017) fundamental research project belonging to "973" programs, named Coordinated Device-to-Device Local Area Networks, i.e. D2D-LAN. This project is organized by Prof. Lingyang Song, in cooperation with Prof. Xiang Cheng, both from Peking University; Prof. Minghua Chen and Prof. Yingjun Zhang, both from Chinese University of Hong Kong; and Prof. Shengli Zhang from Shenzhen University.

Device-to-Device (D2D) communication has been recognized as an efficient way to improve system performance for future wireless networks. In China D2D-related research has attracted a great deal of attention from numerous researchers and wireless engineers in both academia and industry. In the universities, research topics regarding D2D communication cover a wide range, from the physical layer to the MAC layer, and the upper layer, etc. In industry, people mainly look at the possibilities of standardization in 3GPP, as well as real implementation and prototypes, and their current focus is on neighborhood discovery and public safety applications. In almost every local wireless communication conference and seminar, you will see a few presentations about D2D communication.

A large number of research works and projects on D2D systems focus on one-hop (one D2D pair) communication. On the other hand, multihop communications arise in many emerging applications, such as data communication in hotspots. The corresponding research is highly associated with specific applications, such as mobile social networks for advertisement push, and community networks for fast data dissemination. Most of these problems still remain open and are in need of extensive investigation. The D2D-LAN project focuses on these multi-hop application scenarios, catering to the demands of high-efficiency cellular technology, and carries out research from the two perspectives of basic theory and key technology, aiming to develop a new network structure that highly improves spectrum efficiency and system capacity.



A representative study scenario, consisting of small cell, conventional cellular communication, one-hop D2D direct transmission, and D2D-LAN for group communication.

The researchers in the D2D-LAN project believe that there are many challenges different from those faced with traditional wireless networks. One of the major challenges by enabling D2D-LAN communication is to realize efficient data spreading in the D2D network without causing severe disturbance of the original cellular networks. Other challenges to be extensively researched include: identification of services for which D2D communication is useful; radio resource allocation and resource management for D2D links; self-organizing D2D links; and capacity and performance evaluation. Finally, many applications, such as mobile social networks, vehicular ad-hoc networks, or even machine-type communications, will be studied by considering specific constraints.

By now, the project has made innovative progress in various fields of research. The achievements have been published in international journals, conference proceedings, and applied for patents and proposals both in China and aboard. Moreover, the international and Chinese research institutions have given recognition and praise for the work. The results obtained so far through intensive collaboration among the project partners are rather encouraging in comparison with relevant state-of-the-art approaches and thus pave the way to further study of more composite protocols in the future.

For more information, visit the website at **http://wireless.pku.edu.cn/home/songly/973project/home.html**

MEMBERSHIP PROGRAMS/*Continued from page 1*

this situation. One of them was to contact the organizers of technical activities, searching in the official IEEE L31 Report database, and ask them for articles. This was recently implemented a few months ago and we expect to evaluate results next year. We are also managing to have access to the reports that Distinguished Lecturers provide after tours when visiting our region. No doubt that this issue has been the most difficult to address by me.

Stefano: I am confident you will be successful to improve the situation. You also mentioned that you wish to improve the transition process from the Regional Director, whose term is expiring, to the incoming Director.

Pedro: This is a new aspect that I am raising as a problem for our Region. Indeed, the ComSoc policy establishes that RDs are appointed by the President of Comsoc, for two-year terms. In my case, when I assumed this position everything was completely new for me. Despite the fact that I had been a member of the ComSoc Latin America Regional Board (LAB) for six years as, in the region we never had a meeting of the Board. I knew little or nothing about the guidelines and plans of the previous directors. In practical terms, I had to start from scratch.

An interesting discussion has been started with you, Koichi Asatani, and the other regional directors on this subject. This has been very helpful. I have discovered that this problem is not exclusive to Latin America. We have taken some good ideas from the excellent work and organization of the AP Region that I hope to implement in our region. For instance, the AP Region nominates the new AP Director candidate (remember that regional directors are appointed by the President of ComSoc) from a list of active AP Board members who have attended half or more of the last two year's AP Board meetings.

Stefano: I agree that Regional Boards should meet in addition to working by email. How would you proceed to make the LA Board more efficient?

Pedro: Related to the first point, I believe that we need to have more LA Board meetings. Unfortunately, the ComSoc budget does not allow in-person meetings, so we are using virtual-meeting tools. We had our first virtual meeting last November. With the exception of some minor Internet connection problems, everything was fine. I think we should also add more members to the LA Board. Finding volunteers for the LA Board has not been easy. There are several positions within the Board that have not been filled so far. First, this generates overwork for the rest of the Board. Also, this decreases the possibility of finding suitable candidates to be the next LA Director.

DISTINGUISHED LECTURER TOUR/*Continued from page 2*

OpenFlow? A layer-2 protocol, Link Layer Discovery Protocol (LLDP), enables switches to broadcast themselves and identify their neighbor switches. Then through OpenFlow 'Hello' messages, switches report themselves and connectivity to their controller. The controller takes the collected routing information to construct the topology for path computation and then, through OpenFlow 'Modify' messages, configure flow tables at switches.

Where should traffic classification happen? If only TCP/IP-layer traffic classification is needed, it can be done at switches because the packet headers checked by flow tables are TCP/IP headers. But if the application header or even the payload is checked in doing traffic classification, it should be redirected to the extended data-plane, i.e. network function virtualized (NFV) modules. However, for the first packet, redirection to the controller for service chaining (SC) is needed to identify where the NFV modules are.

What about the security of SDN vs. the security by SDN? When talking SDN security, most researchers now talk about securing SDN, especially its centralized controllers, which could be the single points of failures. However, if we view security as a valuable service that requires resources, the operators could offer SaaS (Security as a Service) to their enterprise, residential, and cellular subscribers. Thus, another stream of research should be focused on how to offer SaaS on top of NaaS (Networking as a Service) by SDN.

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

THOMAS MICHAEL BOHNERT (THOMAS.BOHNERT@ZHAW.CH)
EWELL TAN, SINGAPORE (EWELL.TAN@IEEE.ORG)

A publication of the
IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

MCN WORKSHOP/*Continued from page 3*

(EPC) as a Service, based on Fokus OpenEPC, OpenSDNCore, Openstack, and Zabbix, was presented by Marius Corici (Fraunhofer FOKUS, Germany).

After the presentations, a panel including many of the speakers and chaired by Luis M. Correia (IST – University of Lisbon, Portugal) discussed several topics raised during the day, as well as questions from the audience. All presentations are available at the MCN website (**http://www.mobile-cloud-networking.eu/site**) as well as at the IEEE ComSoc Portugal Chapter website (**http://chapters.comsoc.org/Portugal**).

CALL FOR CONTRIBUTIONS
GLOBAL COMMUNICATIONS NEWSLETTER

The Global Communications Newsletter (GCN) appears monthly within *IEEE Communications Magazine.*

It provides an excellent opportunity to present news and events related to communications around the world, as well as activities carried out by IEEE Communications Society chapters in greater detail.

In general, articles published in the Global Communications Newsletter are not technical papers or even technical surveys. Rather, they are short articles informing the IEEE Communications Society community about various activities being carried out and organized in the four corners of the world by the many volunteers who are the true engine of the IEEE Communications Society. Also, major news from the regional telecommunications industry, operators, and academia may be of great interest to our global community.

The relevance, timeliness, and interest of reports published in our Newsletter depend on your cooperation. The willingness of everyone to contribute timely and informative reports is essential to ensure the success of our Newsletter. We look forward to receiving your submissions.

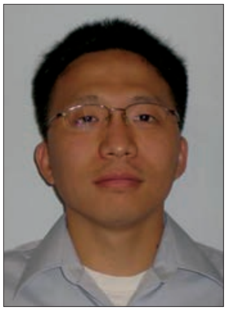
Please submit your articles to the GCN Editor Stefano Bregni (bregni@elet.polimi.it). Submissions should be prepared in standard MS Word DOC format. Articles should begin with title, authors and affiliation, should have length 300 to 1000 words, and may also include 1 or 2 figures.

AMBIENT ASSISTED LIVING COMMUNICATIONS

Joel J. P. C. Rodrigues

Sudip Misra

Haohong Wang



Zuqing Zhu

Nowadays taking care of elderly people and the disabled has become a very important but challenging task. Although the elderly have wisdom and wealth gathered from their life experiences, they require special assistance, higher health insurance costs, and even constant monitoring. By utilizing information process and communications technology, ambient assisted living (AAL) communications open up a new way to address such needs for the aged and the sick. Specifically, AAL uses ambient technologies, including data sensing, processing, transmission, and artificial intelligence, to enable new products, services, and processes that help to provide safe, healthy lives for the aged and disabled. It also supports improved social connections and accessibility to the external world. With the growth of AAL environment, the accessibility gets more challenging for the complex data structure.

In a typical AAL system, ubiquitous computing and sensing integrates microprocessors with common everyday objects, and inter-object communications are enabled via wireless and ad hoc networking. The artificial intelligence empowered by the cloud significantly improves system efficiency. Presently, AAL communications technologies are expected to be transferred first to industry and then to commercial markets. This vision has motivated a voluminous amount of research activities in the field. This Feature Topic intends to capture and expose these activities to the *IEEE Communications Magazine* readership. Through an open call for papers, we received 36 submissions. Eight papers were accepted for final publication after two rounds of highly competitive reviews. The final papers were selected on the basis of originality and significance of the technical work, as well as relevance to the theme topic.

The first article, "A Smart Communication Architecture for Ambient Assisted Living" by J. Lloret, A. Canovas, S. Sendra, and L. Parra, presents an intelligent communication architecture for AAL. It uses artificial intelligence to process the information gathered from several types of communication (e.g., wireless sensor networks, wireless ad hoc networks, wireless mesh networks) over any type of communication technologies (e.g., device to device, machine to machine, sensor-actuator), to know what is happening in the network and detect whether the elderly need to be assisted.

I. Bisio, F. Lavagetto, M. Marchese, and A. Sciarrone contribute an article describing a specific solution called smartphone-centric architecture where smartphones are employed not only as sinks of the health information but also as sensing, processing and transmitting devices. This article is "Smartphone-Centric Ambient Assisted Living Platform for Patients Suffering from Co-Morbidities Monitoring" addresses the case of co-morbidities that indicates the need to acquire a heterogeneous set of data from patients and from their environment. The article also focuses on the information processing capabilities of the smartphone-centric platform.

The article "Overcoming the Challenge of Variety: From Metric-Space Indexing to Big Data Abstraction, the Next Evolution of Data Management for Ambient Assisted Living Communication Systems" by R. Mao, H. Xu, Y. Li, and M. Lu deals with the concept of big data abstraction, using the metric space as the universal abstraction for AAL data types. They show that the metric space is more universal than the conventional multidimensional space and can cover most AAL data types effectively.

The fourth article, "Bayesian Coalition Game for the Internet of Things: An Ambient Intelligence-Based Evaluation," is presented by N. Kumar, N. Chilamkurti, and S. C. Misra. It analyzes a performance evaluation of the Bayesian coalition game among these objects in an IoT environment by using the concepts of game theory and learning automata (LA). In comparison to the available solutions, LA are assumed to be the players, having variable learning rates in the coalition game.

In "LDPA: A Local Data Processing Architecture in Ambient Assisted Living Communications," K. Wang, Y. Shao, L. Shu, G. Han, and C. Zhu present an LDPA on a local server to analyze collected data from ambient sensors. They demonstrate that LDAP disperses the stress of remote centralized processing and data storage, and decreases the workload of a remote health care provider.

Meanwhile, their results indicate that the network load can be reduced and the processing speed improved.

The article “Asynchronous Flow Scheduling for Green Ambient Assisted Living Communications” by D. Wu, Y. Cai, and M. Guizani designs a simple but efficient asynchronous flow scheduling scheme aiming to sense, predict, and realize the AAL applications. They come up with a scheduling architecture by analyzing various device characteristics and user activities, and they also classify the corresponding applications from the aspect of user needs. Their results show that the proposed asynchronous flow scheduling scheme can take energy efficiency into account.

In “Authentication Protocol for Ambient Assisted Living System,” He and Zeadally propose a secure, robust, and efficient authentication protocol for ambient assisted living. They present a detailed security analysis of how their proposed protocol meets the various key security requirements (mutual authentication, anonymity, forward secrecy, etc.) for an AAL system based on intra, inter, and beyond body area networks. Finally, they analyze the computational costs of their authentication protocol, compare its performance results with two recently proposed authentication protocols, and demonstrate the improved performance obtained with their proposed protocol.

The last selected article, “Reliable MAC Design for Ambient Assisted Living: Moving the Coordination to the Cloud,” is motivated by the recent advances in cloud computing. The authors, E. Kartsakli, A. Antonopoulos, A. S. Lalos, S. Tennina, M. Renzo, L. Alonso, and C. Verikoukis, studied the possibility of transfering the network coordination to the cloud while maintaining the data exchange and storage at a local data plane. Then they designed a general framework for the development of cloud-assisted protocols for AAL applications, and proposed a high-performance and error-resilient MAC scheme with cloud capabilities.

To conclude, we would like to express our heartfelt gratitude to 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 Joseph Milizzo, Assistant Publisher, as well as all the other IEEE Communications Society publications staff. We also thank all the authors who have contributed with their strong articles to the success of this Feature Topic, and all the reviewers that did a professional and timely job of reviewing the papers carefully and offering us the opportunity to publish very high-level articles on the timely topic of ambient assisted living communications.

BIOGRAPHIES

JOEL J. P. C. RODRIGUES [S’01, M’06, SM’06] (joeljr@ieee.org) is a professor in the Department of Informatics of the University of Beira Interior, Covilhã, Portugal, and a researcher at the Instituto de Telecomunicações, Portugal. He received his Habilitation in computer science and engineering from the University of Haute Alsace, France, and a Ph.D. degree in informatics engineering, an M.Sc. degree from the University of Beira Interior, and a five-year B.Sc. degree (licentiate) in informatics engineering from the University of Coimbra, Portugal. His main research interests include sensor networks, e-health, e-learning, vehicular delay-tolerant networks, and mobile and ubiquitous computing. He is the leader of the NetGNA Research Group (http://netgna.it.ubi.pt), Chair of the IEEE ComSoc Technical Committee on eHealth, Past Chair of the IEEE ComSoc Technical Committee on Communications Software, a Steering Committee member of the IEEE Life Sciences Technical Community, a Member Representative of the IEEE Communications Society on the IEEE Biometrics Council, and an officer of the IEEE 1907.1 standard. He is the Editor-in-Chief of the *International Journal on E-Health and Medical Communications*, *Recent Advances on Communications and Networking Technology*, and the *Journal of Multimedia Information Systems*, and an Editorial Board member of several journals. He has been general chair and TPC Chair of many international conferences, including IEEE ICC and GLOBECOM. He is a member of many international TPCs and has participated in organizing several international conferences. He has authored or co-authored over 400 papers in refereed international journals and conferences, a book, and three patents. He had been awarded several Outstanding Leadership and Outstanding Service Awards by IEEE Communications Society and several best papers awards. He is a licensed professional engineer (as Senior Member), a member of the Internet Society, an IARIA Fellow, and a Senior Member of IT and ACM.

SUDIP MISRA is an associate professor in the School of Information Technology at the Indian Institute of Technology Kharagpur. He received his Ph.D. degree in computer science from Carleton University, Ottawa, Canada. His current research interests include algorithm design for emerging communication networks. He is the author of over 200 scholarly research papers. He has won eight research paper awards at different conferences. He was awarded the IEEE ComSoc Asia Pacific Outstanding Young Researcher Award at IEEE GLOBECOM ’12. He has also received several academic awards and fellowships such as the Young Scientist Award (National Academy of Sciences, India), Young Systems Scientist Award (Systems Society of India), Young Engineers Award (Institution of Engineers, India), (Canadian) Governor General’s Academic Gold Medal at Carleton University, University Outstanding Graduate Student Award at the Doctoral Level at Carleton University, and the National Academy of Sciences, India Swarna Jayanti Puraskar (Golden Jubilee Award). He was awarded the Canadian Government’s prestigious NSERC Post Doctoral Fellowship and the Humboldt Research Fellowship in Germany. He is Editor-in-Chief of the *International Journal of Communication Networks and Distributed Systems*, Interscience, Switzerland. He has also served as an Associate Editor of the *Telecommunication Systems Journal* (Springer SBM), *Security and Communication Networks Journal* (Wiley), *International Journal of Communication Systems* (Wiley), and the *EURASIP Journal of Wireless Communications and Networking*. He is also an Editor/Editorial Board member/Editorial Review Board member of *IET Networks* and *IET Wireless Sensor Systems*. He has eight books published by Springer, Wiley, and World Scientific. He has been invited to chair several international conference/workshop programs and sessions, and also to deliver keynote/invited lectures in over 30 international conferences in the United, Canada, Europe, Asia and Africa.

HAOHONG WANG [M’04, SM’13] is the general manager of TCL Research America, San Jose, California. He is an inventor of 70+ patents and pending applications, and a co-author of five books and 50+ articles. He is Editor-in-Chief of the *Journal of Communications*, and co-chairs the IEEE Technical Committee on Human Perception and Multimedia Computing. He chaired IEEE GLOBECOM 2010, ICME 2011, and VCIP 2014. He is the recipient of the IEEE MMTC Distinguished Service Award.

ZUQING ZHU [SM’12] (zqzhu@ieee.org) received his Ph.D. degree from the Department of Electrical and Computer Engineering, University of California, Davis, in 2007. From July 2007 to January 2011, he worked in the Service Provider Technology Group of Cisco Systems, San Jose, California, as a senior R&D engineer. In January 2011, he joined the University of Science and Technology of China, where he is currently an associate professor. His research interests are next-generation Internet architecture and software-defined networks. He is a Senior Member of OSA.

A Smart Communication Architecture for Ambient Assisted Living

Jaime Lloret, Alejandro Canovas, Sandra Sendra, and Lorena Parra

The authors are with Universidad Politécnica de Valencia.

In addition, Sandra Sendra is with Instituto de Telecomunicações

ABSTRACT

Intelligent systems and communication technologies have experienced huge advances in the last few years. AAL can benefit from mixing both research fields. This article presents an intelligent communication architecture for AAL. It uses artificial intelligence to process the information gathered from several types of communication (e.g., wireless sensor networks, wireless ad hoc networks, wireless mesh networks) over any type of communication technologies (e.g., device-to-device, machine-to-machine, sensor-actuator), know what is happening in the network, and detect if elderly people need assistance. The article shows the main intelligent algorithms included in the AAL system and the developed software application. Several real measurements validate the operation of our proposal.

INTRODUCTION

Ambient assisted living (AAL) is a term that appeared for the first time in the European Framework Program for research funding. AAL systems pretend to improve the quality of life of special groups of people, including the elderly. It uses information and communication technologies to develop applications and services for elderly people in order to help them in their daily affairs. It can allow them to stay alone at home longer and be more independent, and reduce the time with caregivers. Sensors can be used to increase the safety of their lifestyles and in their home environment [1]. The importance of AAL for elderly people is growing because recent changes in the lifestyle of developed countries and improvements in the medicine field are increasing the size of the older population. In Europe, it is expected that in 50 years, 30 percent of the population will be over 65 years old; at present, this is only 17 percent [2].

Elderly people living alone have more accidents. For example, in the district of Kaiserslautern, Germany, 30 percent of people over 65 who are living alone at home suffer at least one fall per year. Fifty percent of them suffer several falls, and 20 percent of them suffer recurring falls over only six months. Without help those people cannot get up by themselves and may remain on the floor for long periods of time. Many complications may derive from these falls such as decubitus ulcers, hypothermia, pneumonia, or even death. Personal emergency response systems use emergency push buttons that can be inappropriate to solve some of these situations. A fallen person might not be able to get to the button or may be unconscious, in shock, or nervous, thus forgetting that the systems are there or how to use them [3, 4]. Systems that do not require the attention of elderly people are needed. Systems like AAL that use sensors, which can prevent those dangerous situations, and actuators such as alarms, phone calls, or SMS, which can be sent to advise if a dangerous situation occurs, can provide much more efficient and usable care for elderly people.

When implementing an AAL, several features should be considered [5]. On one hand, sensors should be non-invasive systems. They can be embedded in clothes, shoes, watches, or glasses. Thus, people will not mind wearing them. If sensors were visible, users could be discriminated against by other persons. In addition, sensors must have wireless communication interfaces to let people move away from their homes. Being alone at home for long periods of time can cause depression or other psychological problems. On the other hand, AAL systems should be able to scan the local environment to obtain useful information and exchange information with similar nodes in their neighborhood. AAL systems should be able to act when they detect any abnormal situation without the explicit request of a user. This is very important in cases when a person is unconscious after a fall. Finally, AAL systems should be able to adapt by themselves when abnormal situations occur. These systems may depend on the special needs of each person. They should also consider special methods to interact with users, such as voice orders or gestures instead of keyboards or mice.

Nowadays, it is easy to find lots of AAL proposals based on several sensors to measure weight, blood pressure, glucose, oxygen, temperature, location, and position. Each system is deployed using a communication technology such as ZigBee, Bluetooth, ZWave, USB, and Ethernet, among others. In addition, the most used interfaces are developed for tablets and

0163-6804/15/$25.00 © 2015 IEEE

smartphones, although we can also find applications for health systems or set-top boxes. Generally, these kinds of systems are focused on solving several issues in services such as healthcare provision, disease management, diet and fitness, personal health records, and person location [6].

The main goals AAL systems must accomplish are to ensure a person's welfare, that is, monitor illnesses, control the provision of fresh food, generate alerts for medical personnel in case of falls or abnormal situations, monitor medication reminders, and enhance social relations by enabling people-to-people communications. Recent advances in household appliance monitoring are allowing all of the above [7].

The sensors used in AAL systems should be cheap and need low maintenance. AAL systems must be cheaper than having a caregiver. However, although AAL systems can be very useful, this technology cannot replace human support. They should be considered as complementary to human assistance. AAL systems can offer a good way to assist elderly people in tasks such as measuring physical parameters and vital signs.

Although most of the applications are focused on elderly people, there are other groups with special needs such as the visually and hearing impaired, and people with physical disabilities or chronic diseases. Babies and children can also be considered within this special group. The main environments where those systems are developed are homes, hospitals, and nursing homes [1]. Nowadays, there are still no commercial solutions at affordable prices to reach anyone. However, there are many research projects, such as AMIGO, GENESYS, MPOWER, OASIS, PERSONA, and SOPRANO [6], which are working to make it possible.

In this article, we present a communication architecture for AAL that uses artificial intelligence with the data gathered from the sensors and the detected behavior of people's movement, traffic patterns, and network frame operation results. It is able to react in strange situations or when the behavior of elderly people is not what is expected. Several artificial intelligence systems have been included in the decision algorithms in order to provide the most appropriate actions for the proposed AAL architecture.

The rest of this article is structured as follows. The following section shows the main proposals and investigations in AAL field. Then the proposed system and its procedure are shown. After that we show several implemented cases and the system deployment. Then the verification of our system is provided. The final section draws the conclusion and suggests future work.

Related Work

Within the AAL field, we can find several systems, proposals, and investigations. Most of them present simple definitions and ideas. Very few present real implementations, testbench implementations, or even simulations.

In general, AAL systems are developed to offer better quality of life to elderly and disabled people. As M. Memon *et al*. show in [8], the main investigated issues related to AAL are: AAL system architectures, design and development methodologies for AAL systems and services, technology standards and specifications, and security, privacy, and data protection. Some other studies present several sensors and actuators that can be integrated in AAL systems [9].

There are a wide range of AAL applications and systems proposed to improve the quality of life of elderly people. Many research projects are focused on smart scenarios and systems to monitor an environment for extracting information from them. Considering this, R. Blasco *et al*. [10] present the design, implementation, and assessment of a Smart Kitchen which creates a smart environment that increases elderly and disabled people's autonomy in their kitchen-related activities. The system is based on a modular architecture that integrates systems such as household appliances, sensors, and user interfaces as well as associated communication standards and media (power line, radio frequency, infrared, and wired). The software is based on the Open Services Gateway initiative (OSGi), which allows complex systems to be designed and can easily be scaled to meet users' needs. The system was evaluated with a large number of real users and careers in two living laboratories. The results demonstrated that the system had adequate functionalities for elderly and disabled people. This system could prolong the time elderly and disabled people remain independent in their own homes.

There is a wide range of AAL applications and systems proposed to improve the quality of life of elderly people. Many researches are focused on smart scenarios and systems to monitor an environment for extracting information from them.

M. Vacher *et al*. [11] present a real-time audio analysis system called the AuditHIS system. This system is devoted to audio analysis in a smart home environment. AuditHIS is a software application developed to ensure online sound and speech recognition. Eight microphones are connected to an acquisition board, and all channels are analyzed simultaneously. The audio is classified by energy thresholds. The sounds are detected and classified as daily life sound or speech and sent to a sound classifier. The sentences and sounds are processed by a developed acoustic and language model. Finally, the system is able to make a record of each audio event and store it in the computer for further analysis. This system can be used for sound and speech processing, activity monitoring, and distress situation recognition.

L. Meinel *et al*., [12] presented an automated video-based real-time surveillance system. The system uses an omnidirectional camera and a multiple object tracking method for applications on the AAL field. The presented system can track multiple persons entering and leaving a room and is able to monitor a complete room with a single camera. For each detected person, the camera generates a virtual perspective using fast transformation techniques. The software is implemented in an embedded platform that acts as a smart sensor.

Ayala *et al.* [13] focused their work on the design and implementation of an agent-based AAL system. The proposal is based on the MalacaTiny agent architecture. MalacaTiny agents use aspect orientation to enable a highly reconfigurable architecture that has self-configuring capacity and is platform-neutral. Agents can be

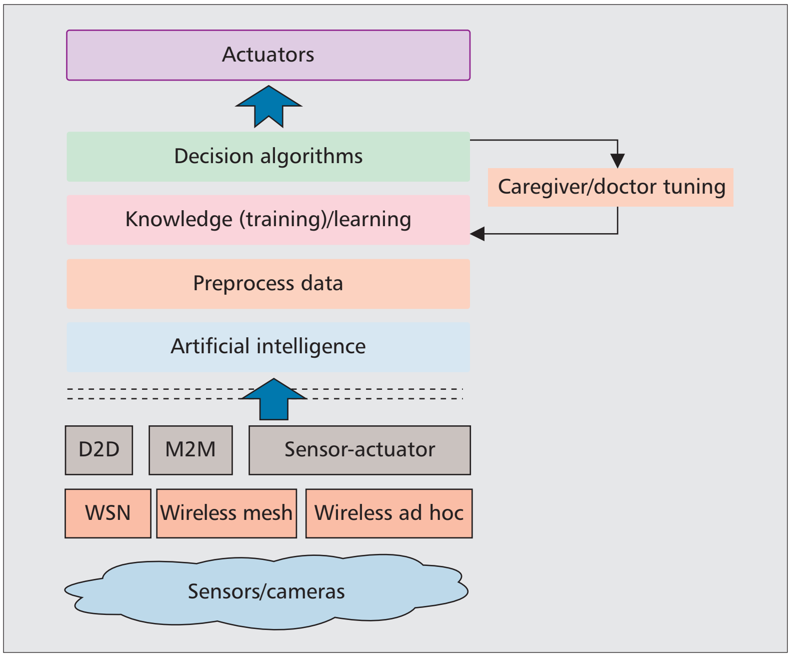


Figure 1. AAL architecture.

executed in different devices through native communication protocols of each device providing autonomic management tasks for helping elderly people. The authors implemented the system in several scenarios. Moreover, they analyzed the system performance in terms of response times of reconfiguration and wireless data exchange. The proposed architecture can be extended with new types of sensors and other kinds of AAL devices.

As we can see, none of the aforementioned articles included an intelligent system that, besides sensor data, takes into account unusual human behavior in order to intelligently decide on the most appropriate action. Moreover, the system is able to detect bottlenecks in the network and balance the number of connections of a single device. This study is motivated by the increasing tendency of society to equip their homes with large numbers of wireless sensors, mobile devices, and actuators, regardless of the fact that a large number of wireless devices may worsen the global wireless network behavior.

Architecture Overview

The proposed system uses the information gathered by the sensors and cameras included in smartphones, the sensors of the wireless sensor network (WSN), and IP cameras and devices located in the wireless network. The communication technology used in our proposal is machine-to-machine (M2M), where both wireless and wired systems communicate with other devices in order to gather enough information and decide whether or not an action or alarm should be activated. The proposed system can also be used in device-to-device (D2D) or sensor-actuator networks.

Based on the information presented above and the requirements of elderly people, we formulate the next model paying attention to the different technologies, application environments, and degrees of dependence of elderly people. The proposed system includes next features:

- Independent of communication technology
- Supports different manufacturers in the same network
- Adaptable to different needs of elderly people
- Suitable for different environments
- Low energy consumption to avoid large charge periods
- Low bandwidth consumption to allow it to run over many wireless technologies
- Easy to use
- Scalable in order allow new sensors to be added or removed when needed
- Low size and not easily detectable in order to avoid visibility
- Allow and control the movement and displacement of elderly people

Mobile devices registered in the system gather information from other devices and sensors in the network. When this information is received by the device, the artificial intelligence module uses the processed information to know what is happening and use it in the decision algorithms. In our case, we have used an artificial intelligence technique based on supervised learning methods (e.g., statistical and neural network methods). Wireless network parameters of all nodes/devices, sensed data, and the video received from IP cameras are the input parameters of each training corpus. The samples included in each corpus depend on the context in which it is going to be applied. For example, there is a corpus to estimate the location of an elderly person inside a house, another to identify pain using face image recognition, another to identify falls, another the detection of anomalies in their behavior, and so on. In order to perform supervised learning, all corpus samples are tagged by the caregiver or doctor. Each tag includes the position and state of the person (normal or anomalous). The position data is automatically included using the wireless network though a self-location system [14]. After the tagging process, preprocessing occurs, where a set of sensor, IP camera, and mobile samples are related to the position and state of the person. Data normalization has been used for an efficient classification process. The learning methods used in the system have been neural networks or statistic, depending on the case; for example, position estimation, falls, and human mobility have used neural networks; almost all the rest have used statistical methods. After the training phase, and having finished the learning phase, a sophisticated series is obtained. In order to determine what should be performed and with which actuator, a decision algorithm and rules including all possible cases have been used. Moreover, the system lets us predict what will happen and if there is an anomaly in a person's behavior. Figure 1 shows the proposed architecture.

In our architecture, every node gathers measurements via two methods. The first one is direct sensing, which obtains the variables sensed from the physical medium (i.e., the information measured physically). The second is indirect sensing, where the measurements are obtained

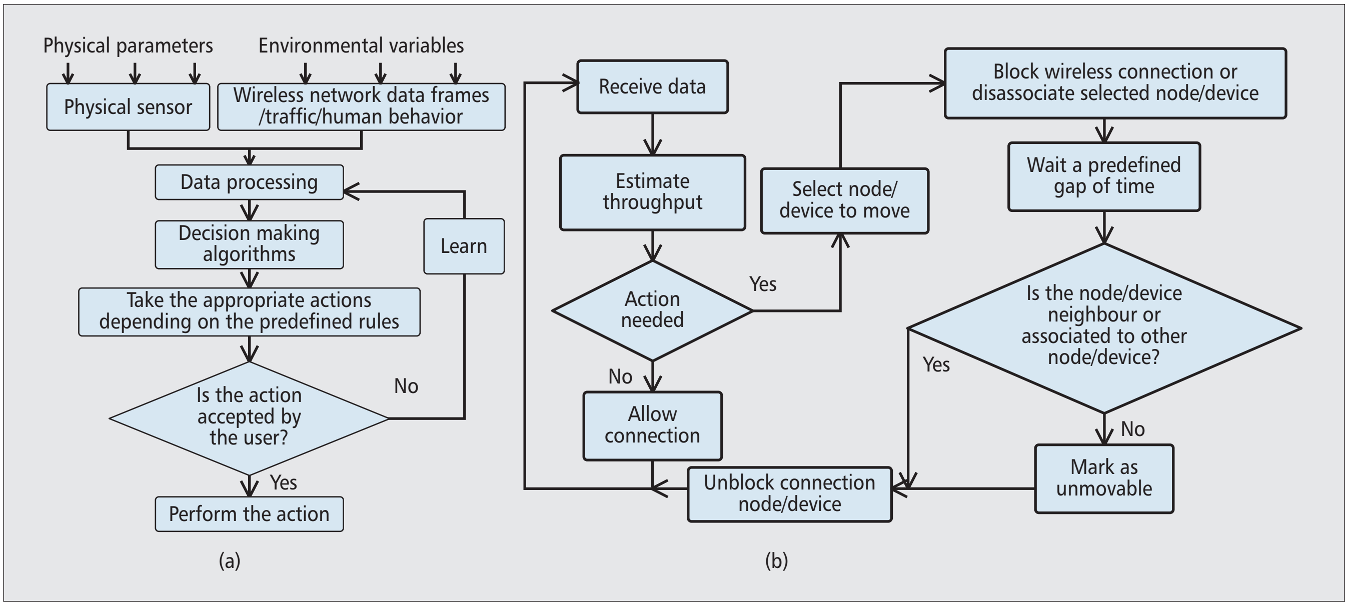

Physical parameters
Environmental variables
Physical sensor
Wireless network data frames /traffic/human behavior
Data processing
Decision making algorithms
Learn
Take the appropriate actions depending on the predefined rules
Is the action accepted by the user?
No
Yes
Perform the action
(a)
Receive data
Estimate throughput
Action needed
Yes
No
Allow connection
Select node/ device to move
Block wireless connection or disassociate selected node/device
Wait a predefined gap of time
Is the node/device neighbour or associated to other node/device?
Yes
No
Mark as unmovable
Unblock connection node/device
(b)

Figure 2. Architecture operation: a) data process flow; b) procedure algorithm when the number of wireless devices increases.

from the parameters of the wireless network (received signal strength, lost frames, disconnections, etc.), the amount of data traffic during a period of time, and the location of the detected human in order to use them in the decision algorithms. Moreover, they can be used to extrapolate some conclusions that can be useful in attaining a complete picture of what is happening in the environment and obtaining the correct prediction.

In order to achieve our goal, we have studied how some data network parameters are affected by the environment in certain cases. Moreover, we have collected the most useful parameters for the environmental monitoring and their relationship to what is happening in the environment. They are mainly related to human behavior and habits. When a sensor node, device, or camera creates an alert (e.g., because it has reached a threshold or a rare sensed value is obtained), the sensed variables and environmental parameters are also sent to the mobile of the affected person. Since the information is sent only to a device that belongs to the network (rather than to all the devices of the network), we avoid unnecessary message forwarding and additional overheads, thus saving energy.

Figure 2a shows the data process flow. Initially, data is gathered by the physical sensors and indirectly from the wireless network data frames, traffic, and human movement or actions. These data are processed and used by the decision algorithms in order to take the appropriate actions based on the rules created for the AAL. When an action (e.g., sending an alarm) is going to be performed, the application waits for a short period of time for user intervention in order to cancel it. If the alarm is not cancelled, it is sent; otherwise, the decision is learned as a false positive and taken into account for future actions.

We have also included decision algorithms that are used to improve network performance. They use parameters such as sensor/device placement inside the network, network traffic, and sensor/device movement. They allow balancing the amount of wireless links in sensor/device nodes, especially highly dense networks, balancing the traffic load when too much data is transmitted through the network (sensed data, video data, sound data), and predicting resource reservation (very useful when nodes/devices are moving in order to provide resources before they arrive at the next place). Figure 2b shows an algorithm that lets nodes/devices balance the number of simultaneous wireless connections to avoid decreasing network performance and balance the wireless connections between nodes/devices. This algorithm is focused on allowing more wireless sensors, devices, and actuators in homes without having an impact on the global wireless network behavior. In the algorithm, each node analyzes the connection/association requests. When the node/device receives a new request, it estimates the available throughput and decides if it can allow a new connection; if not, the connection of this device is blocked in order to let it choose another node/device with which to connect. When it has verified that it has joined the network through another neighbor, the block is removed.

In our architecture, redundancy and fault tolerance can easily be added just by including more smartphones, sensors, IP cameras, and devices located in the wireless network. Moreover, a central server can also be added in order to receive all decisions and alarms sent between devices. A central server can also help detect node failures and loss of information.

Developed AAL Application

The proposal can be seen as a group-based system, where each group is formed by a network with the same types of nodes (WSN, wireless mesh network, or wireless ad hoc network). A group-based architecture provides more benefits than regular architectures [15]. They provide

In our architecture, redundancy and fault tolerance can easily be added just by including more smart phones, sensors, IP cameras, and devices located in the wireless network. Moreover, a central server can also be added in order to receive all decisions and alarms sent between devices.

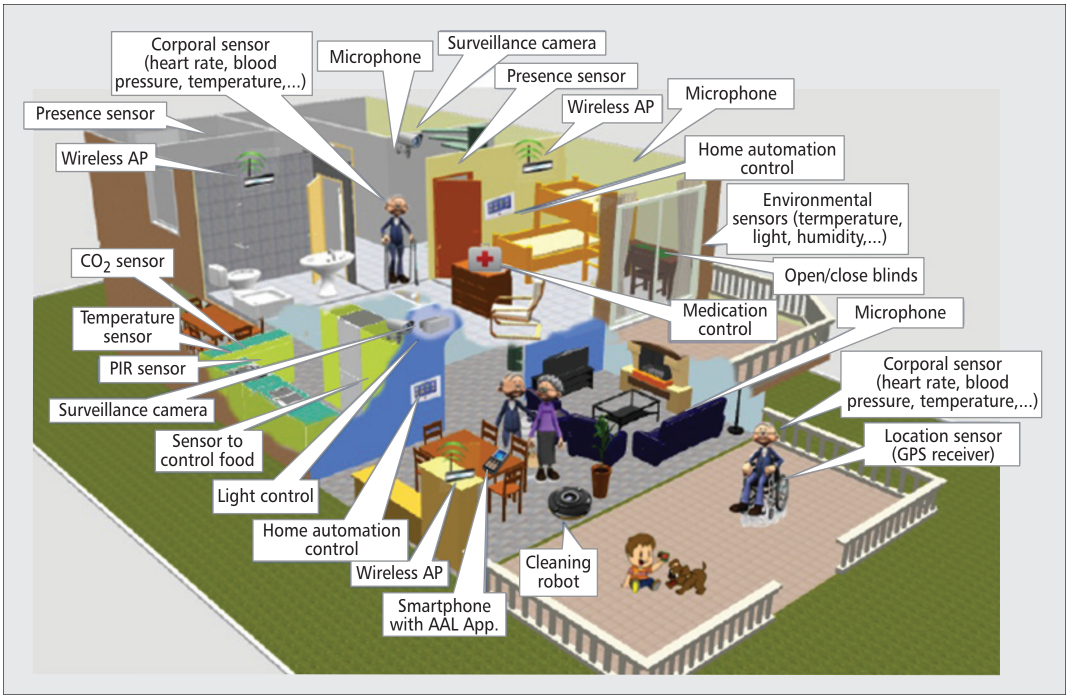


Figure 3. Regular AAL elderly people's residence or house.

more scalability, higher productivity, lighter network load, more energy saving, lower communication costs, and so on.

Using the developed application, computing devices(e.g., smartphones, tablets, laptops) will receive the data corresponding to the user handling the device. Since the communication technology is M2M or D2D, mobiles have end-to-end connections with sensors/devices/actuators to which they will be related to by which they will be affected. Then the mobile device uses the implemented artificial intelligence technique and decision algorithms to take the appropriate actions or send the required alarm.

Figure 3 shows a regular AAL residence or house. The house is full of sensors, such as a presence sensor, CO$_2$ sensor, and temperature sensor, and actuators, such as light control, home automation control, medication control, and even a cleaning robot.

Next, there is a list with some cases that are included our AAL proposal (this list provides a picture of what it is able to do; it can be extended with more cases):

- When microphones (e.g., the microphones embedded in IP cameras) detect a shout, because its sound level is over a threshold, the device recognizes the voice of the shouting person and applies a pattern recognition system in order to know if this person habitually speaks loudly or it is common behavior. When no pattern is found, an alarm is produced.
- When the accelerometer and gyroscope of a wearing device has detected a fall, or an image processing system, using the video obtained from a camera, has detected a fall, an alarm is sent.
- When sensors or devices (presence sensors, mobile phone sensors, cameras, etc.), detect that there is no human movement when there is usually some activity, an alarm is activated.
- When a body sensor registers a value that is higher or lower than a threshold, an alarm is activated.
- When a person reaches a room that is dark, the light in that room is switched on.
- When a person is seated on a sofa or arm chair that is generally used to watch TV, the TV is switched on.
- When the person arrives at home and opens the main door, the light of the hall of the house is switched on.
- When there is no light outside because it is a dark afternoon or night has come, the blinds should be lowered. The opposite should happen when the light rises in the morning.
- When sunlight directly hits a window or glass, an awning is lowered.
- When the face recognition system identifies pain, an alarm is sent.

An example is given in Fig. 4a. The screen shows the day activity prediction graph. In this case, the elderly person being monitored has arrived at the living room; since there is no light (because of a light sensor and the time of day), the living room light is automatically switched on. There are several implemented cases where alarms are produced. Figure 4b shows an alarm produced because the system has detected that the blood pleasure is high (although it has not reached a threshold), and the system has detected an unexpected body movement (it can be measured by the gyroscope and the accelerometer of the device worn by the elderly person, or by image recognition). If the user cancels the alarm, it has been a false positive, so the system will tag it as such.

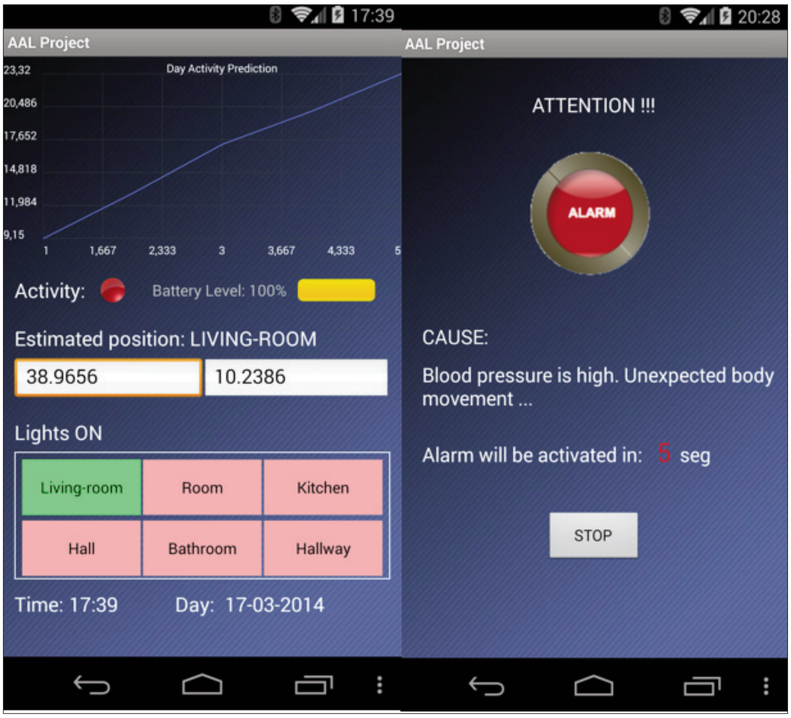
System Verification

In order to verify the proposed system, we have performed several tests in different case studies.

In the first test, a training phase shows how the mobile learns the user's habits. The system uses the accelerometer values (X, Y, Z) and three variables: GPS position, time, and battery power. The learning process is as follows. The mobile device gathers samples from the accelerometer sensor when it detects movement. The movement is detected by variation in the accelerometer values. After gathering enough samples, the application software seeks a behavioral pattern taking into account the hours of the day. It also seeks to find in which hour(s) of the day there is more movement and when there are repeated movements. The system will also try to discard dispersed values. During a period of inactivity, the mobile device also gathers data, but it is tagged as mobile inactive. Figure 5a shows the mobile device training and learning process example. It shows the samples obtained inside the house with a full battery. We can see that the activity test predicts a high level of accuracy in the patterns obtained during the learning process. In Fig. 5b, we also take into account the battery value. We performed the same process as before, but this time the battery power values are also included in the learning process. After gathering measurements for all cases, the amount of battery power is also related to human behavior. This seems to be because it has more connections and disconnections with the wireless nodes and because movement makes the device take more measurements from the sensors, which implies more power consumption. Moreover, this relationship can also occur because of the person's behavior. A person can be aware of the battery state and decide not to move away because there is low battery and the mobile must be charged.

In the second test we analyze the algorithm procedure used to balance the number of connections of a sensor/device (Fig. 2b). In order to perform the test we used two Cisco Aironet access points (APs) series 1100 and a smartphone, Samsung S4. The smartphone is associated with the first AP and sends the sound gathered by the microphone. Suddenly, the system decides to relocate the smartphone to another AP, so it is blocked by the first AP and re-associated with the second AP.

Figure 6a shows the number of bytes per second during 40 s, where the algorithm shown in Fig. 2b is executed. In the 15th second, the device is thrown off the AP. The peak inside the red circle indicates the disassociation process. This peak has a value of approximately 25 kB/s.



Activity: Battery Level: 100%

Estimated position: LIVING-ROOM

38.9656 10.2386

Lights ON

Time: 17:39 Day: 17-03-2014

ATTENTION !!!

CAUSE:

Blood pressure is high. Unexpected body movement ...

Alarm will be activated in: 5 seg

Figure 4. Smart application: a) lights switched on; b) alarm activated.

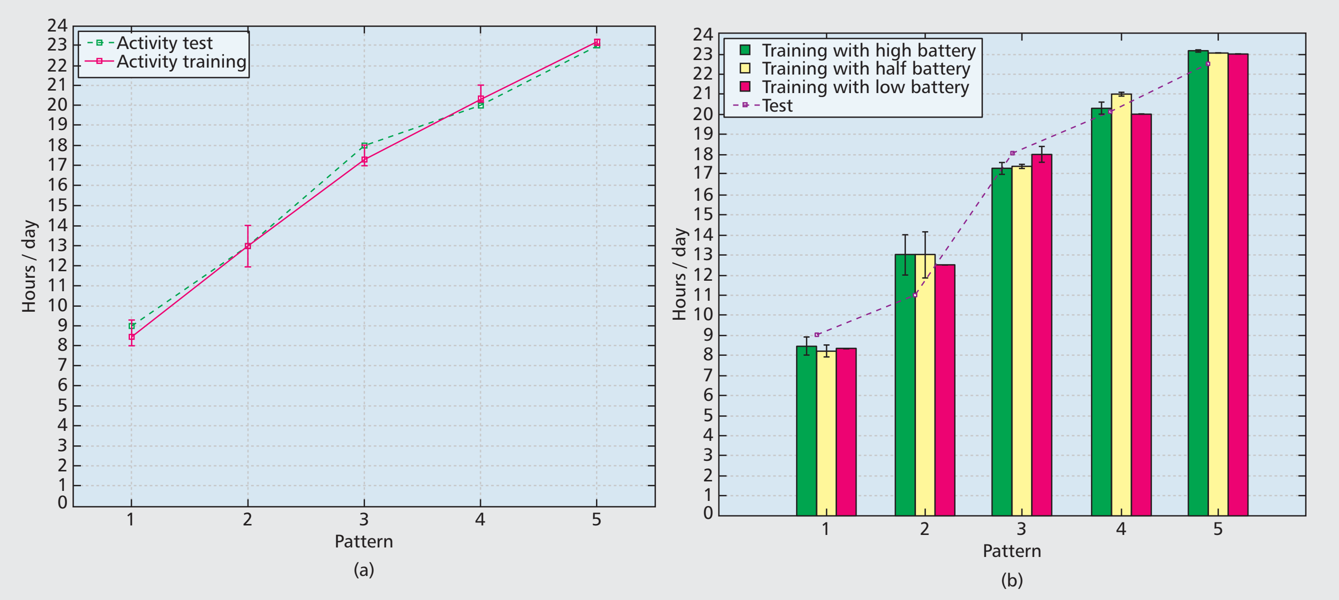


Figure 5. Training process. a) the mobile training pProcess; b) how the battery affects the training process.

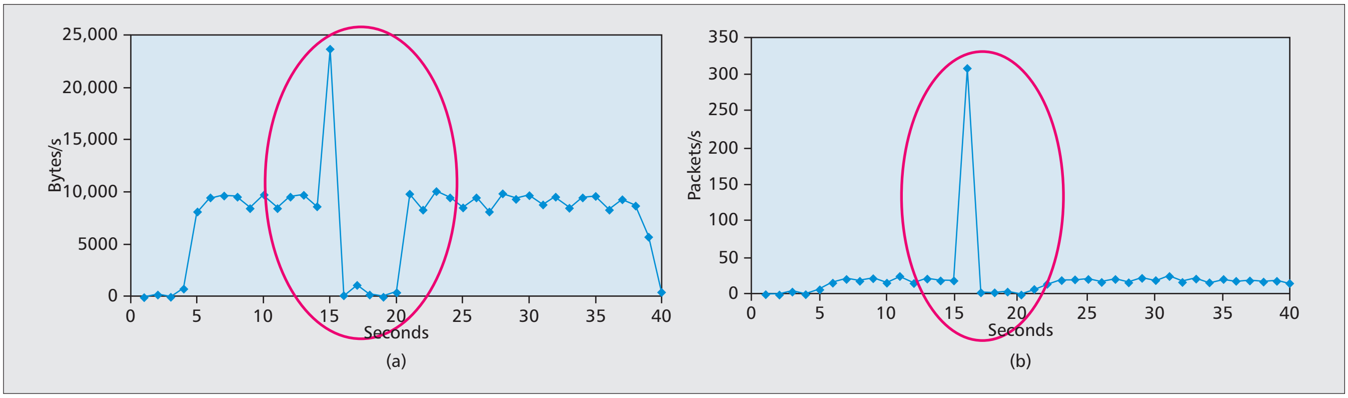


Figure 6. Network performance. a) Bytes/s in the network; b) Packets/s in the network.

Then we observe that there is an interval of 3 s where no data is transmitted. This is due to the re-association process. Figure 6b shows the number of packets per second obtained for the same process. The change of AP implied a peak of around 300 packets/s to the data network (highlighted by the red circle).

Real measurements show that in spite of the amount of bandwidth used by each sensor, actuator, or camera, the performance of a node/device decreases greatly when there are more than 48 wireless connections.

Conclusion

This article has shown the design and development of a smart communication architecture for AAL using a communication technology such as device-to-device, machine-tomachine, or sensor-actuator. It uses the information gathered from sensors, data traffic patterns, and a person's behavior (and habits) in order to make decisions and send alarms. In order to achieve our goal, we have applied a supervised intelligent system that learns from user decisions, and caregivers and/or doctors for future cases. The system presented in this article includes some designed algorithms, representation of the knowledge acquired from the environment, and information management of the wireless network.

In future work, we will include more cases in order to better assist elderly people in their regular living. Moreover, robots will also be incorporated.

References

[1] J. Tomas *et al.*, "Sensors and Their Application for Disabled and Elderly People," *Handbook of Research on Personal Autonomy Technologies and Disability Informatics*, IGI Global, 2011, pp. 311–30.

[2] M. J. O'Grady *et al.*, "Towards Evolutionary Ambient Assisted Living Systems," *J. Ambient Intelligence Humanized Computing*, vol. 1, no. 1, 2010, pp. 15–29.

[3] H. Storf *et al.*, "An Event-Driven Approach to Activity Recognition in Ambient Assisted Living," *Ambient Intelligence*, Springer Berlin Heidelberg, 2009, pp. 123–32.

[4] T. Kleinberger *et al.*, "Ambient Intelligence in Assisted Living: Enable Elderly People to Handle Future Interfaces," *Universal Access in Human-Computer Interaction*, Ambient Interaction, Springer Berlin Heidelberg, 2007, pp. 103–12.

[5] J. Nehmer *et al.*, "Living Assistance Systems: An Ambient Intelligence Approach," *Proc. 28th Int'l. Conf. Software Engineering*, Shanghai, China, 20–28 May 2006, pp. 43–50.

[6] M. H. Tazari, R. Wichert, and T. Norgall, "Towards a Unified Ambient Assisted Living and Personal Health Environment," *Ambient Assisted Living*, Springer Berlin Heidelberg, 2011, pp. 141–55.

[7] J. Lloret *et al.*, "Ubiquitous Monitoring of Electrical Household Appliances," *Sensors*, vol. 12, no. 11, 2012, pp. 15,159–91.

[8] M. Memon *et al.*, "Ambient Assisted Living Healthcare Frameworks, Platforms, Standards, and Quality Attributes," *Sensors*, vol. 14, no. 3, 2014, pp. 4312–41.

[9] S. de Miguel-Bilbao *et al.*, "Short Range Technologies for Ambient Assisted Living Systems in Telemedicine: New Healthcare Environments," *Telemedicine*, Dr. Ramesh Madhavan, Ed., InTech, 2013.

[10] R. Blasco *et al.*, "A Smart Kitchen for Ambient Assisted Living," *Sensors*, vol. 14, no. 1, 2014, pp. 1629–53.

[11] M. Vacher *et al.*, "Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges," *Int'l. J. E-Health and Medical Commun.*, vol. 2, no. 1, 2011, pp. 35–54.

[12] L. Meinel *et al.*, "Automated Real-Time Surveillance for Ambient Assisted Living Using an Omnidirectional Camera," *Proc. 33rd Int'l. Conf. Consumer Electronics*, Las Vegas, NV, Jan. 10–13, 2014, pp. 396–99.

[13] I. Ayala, M. Amor, and L. Fuentes. "Self-Configuring Agents for Ambient Assisted Living Applications," *Personal and Ubiquitous Computing*, vol. 17, no. 6, 2013, pp. 1159–69.

[14] M. Garcia *et al.*, "The Development of Two Systems for Indoor Wireless Sensors Self-location," *Ad Hoc & Sensor Wireless Networks*, vol. 8, nos. 3–4, 2009, pp. 235–58.

[15] J. Lloret *et al.*, "Improving Networks Using Group-Based Topologies," *Computer Commun.*, vol. 31, no. 14, 2008, pp. 3438–50.

Biographies

Jaime Lloret [SM] (jlloret@dcom.upv.es) received his M.Sc. in physics in 1997, his M.Sc. in electronic engineering in 2003, and his Ph.D. in telecommunication engineering (Dr.Ing.) in 2006. He is a Cisco Certified Network Professional Instructor. He has worked as a network designer and administrator in several enterprises. He is currently an associate professor at the Polytechnic University of Valencia. He is head of the Communications and Remote Sensing research group of the Integrated Management Coastal Research Institute and the Active and Collaborative Techniques and Use of Technologic Resources in Education (EITACURTE) Innovation Group. He is director of the University's diploma, Redes y Comunicaciones de Ordenadores, the university's Expert Certificate in Tecnologías Web y Comercio Electrónico, and the university's Master's in digital post production. He is currently Chair of the Internet Technical Committee of IEEE Communications and Internet Societies. He has authored 12 book chapters, and has more than 340 research papers published in national and international conferences and international journals. He has been a Co-Editor of 15 conference proceedings and a Guest Editor of several international books and journals. He is Editor-in-Chief of the international journal *Networks Protocols and Algorithms*, IARIA Journals Board Chair, and an Associate Editor of several international journals. He is currently Chair of the Working Group of IEEE Standard 1907.1. He has been General Chair (or Co-Chair) of 27 International workshops and conferences. He is an IARIA Fellow.

Alejandro Canovas (alcasol@upvnet.upv.es) received his Ph.D. in telecommunications engineering in 2005. He obtained the title of University Specialist in Networks and Communications of Computers in 2007 and that of Master in Artificial Intelligence, Pattern Recognition and Digital Imaging in 2012. He is a Cisco Certified Network Associate Instructor. He worked as a programmer on several projects with the Polytechnic University of Valencia in collaboration with different enterprises. He is currently a Ph.D student at the Polytechnic University of Valencia. He is member of the Communications and Remote Sensing research group of the Integrated Management Coastal Research Institute. He has had more than 40 research papers published in national and international conferences, and international journals (13 with ISI Thomson Impact Factor). He is Assistant Editor of the international journal *Networks Protocols and Algorithms*. He has been involved in several international conferences and workshops as Technical Program Committee Member and Webchair (ACCESS 2010, SCPA 2012, GreeNETS 2012, ACCESS 2012, eL&mL 2012, ACCESS 2012, CONTENT 2013, SSPA 2013, GreeNETS 2014, Chinacomm 2014, eLmL 2014, SSPA 2014, and MARSS 2014). He is currently a member of the Working Group of IEEE Standard 1907.1. He is currently webchair of SCPA 2015 and is involved as a Technical Program Committee member of some international conferences and workshops.

Sandra Sendra (sansenco@posgrado.upv.es) received her degree in technical engineering in telecommunications in 2007. She received her M.Sc. in electronic systems engineering in 2009 and her Ph.D. in electronic engineering (Dr. Ing.) in 2013. She is a Cisco Certified Network Associate Instructor since 2009. She is author or coauthor of more than 60 papers in SCI journals, peer-reviewed conference proceedings, books, and international book chapters. She is Editor-in-Chief of the international journal *WSEAS Transaction on Communications*, Guest Editor for several Special Issues of international journals, and Associate Editor of *Network Protocols and Algorithms*. She has been involved in more than 100 program and organization committees of international conferences. Her research interests include energy saving techniques in electronic circuits, sensor deployment, WSN, UWSN, and the application of these technologies for environmental monitoring.

Lorena Parra (loparbo@doctor.upv.es) received her degree in environmental science in 2012, her M.Sc in environmental assessment and monitoring of marine and coastal ecosystems in 2013, and a second M.Sc. in aquaculture in 2014. She is a Ph.D. student at Polytechnic University of Valencia. Her research interests are focused on integration of new technologies for environmental monitoring, especially in underwater environments. She is an author or coauthor of several papers in SCI journals. She has been involved in several program and organization committees of international conferences.

Smartphone-Centric Ambient Assisted Living Platform for Patients Suffering from Co-Morbidities Monitoring

Igor Bisio, Fabio Lavagetto, Mario Marchese, and Andrea Sciarrone

The authors are with the University of Genoa.

ABSTRACT

Recently, patients suffering from a set of physical and mental limitations, called co-morbidities, are often treated at home. In this environment, modern communication systems represent a great support to implement Ambient Assisted Living platforms aimed at monitoring patients at home because they enable the seamless integration of heterogeneous sensing units, medical devices, and ubiquitous access to data. This article describes a specific smartphone-centric architecture where smartphones are employed not only as hubs of the health information but also as sensing, processing, and transmitting devices. Smartphones have both short-range (Bluetooth and WiFi employed for local information exchange) and long-range (GPRS, 3G/4G, and WiFi employed as Internet access) communication capabilities; information processing capabilities offered by modern platforms often equipped with different CPUs and with flexible and efficient software; and sensing capabilities implemented through sensors embedded into smartphones such as GPS receivers, accelerometers, microphones, and radio interfaces or through external sensors added to smartphones by cables or connected through local radio interfaces. The specific case of co-morbidities considered in this article implies the necessity to acquire a heterogeneous set of data from patients and from their environment. For this reason this article highlights the information processing capabilities of the introduced smartphone-centric platform. Audio, localization, and movement information processing have been evidenced as well as the specific implementations of these capabilities and their performance.

INTRODUCTION

Due to demographic change and increasing healthcare costs, often patients suffering from a given pathology are treated at home. This approach allows continuous monitoring and treatments, enables improvements of the health status, makes patients and their families play an active role in the care process, and reduces healthcare costs related to hospitalization management. However, the transition of treatments formerly conducted in the hospital to home environments is not possible without obstacles, also from communications, networking, and signal processing viewpoints. It is important to focus on the following major challenges:

- Treatment is not constantly supervised and personalized. At home there are no medical experts who monitor the situation of a patient and immediately adapt the prescribed treatment accordingly in case of need.
- Different treatments applied in co-morbid patients (i.e. patients suffering from a set of physical and mental limitations) may contrast with each other. For effective treatment, it has to be considered that co-morbidity is not merely an accumulation of different illnesses. Rather, a patient's condition is determined by the mutual interaction of different diseases. A supervision action carried on within a hospital, but missing at home, can mitigate the problem.
- Best practice may not be carried out or standardized home-treatment. Most medical protocols and guidelines are intended for clinical treatments and are not easily mapped to home treatment. Moreover, often not enough reliable data are available to get statistical validation to develop home-based treatment guidelines.

Information and communications technologies, currently employed in the medical context to increase safety and efficiency and to enable remote patient monitoring, may help tackle these challenges (see [1–9], among many others). Modern communication systems represent a great support in health-related applications and enable the design and implementation of Ambient Assisted Living (AAL) platforms aimed at monitoring patients at home. For example, many smartphone apps evaluating health status have been developed, and the U.S. Food and Drug Administration (FDA) approved the use of smartphones for the collection of medical data in online-databases concerning vital data monitoring services.

Unfortunately, these solutions often lack of

0163-6804/15/$25.00 © 2015 IEEE

interoperability with other devices because they do not implement existing data exchange standards (e.g. HL7) or, as most medical devices for remote monitoring, they are designed as isolated products. This situation hinders their application for co-morbid patients treated at home.

The integration and interoperability of AAL platforms for data exchange could help the development of remote monitoring services for co-morbid patients and of tailored medical surveillance systems.

In this context the acquisition and management of health-related data is a topical task to implement the decentralized treatment of co-morbid patients where multiple medical specialties are involved. Remote assistance requires, on one hand, personalized devices applied to assure continuous information exchange, and on the other hand, ubiquitous access to make feasible an integrated treatment of all involved healthcare providers. This implies the need for a communication architecture able to manage the aforementioned issues by implementing a seamless integration of heterogeneous sensing unit, and medical devices, and by providing ubiquitous access to data. The remainder of this article is organized as follows. The following section introduces the general characteristics of a Communication Architecture for Co-Morbidities Management and describes how the acquired information may be employed by patients, physicians, and medical device manufacturers. We then present a specific implementation of the mentioned architecture based on the employment of smartphones, which are used simultaneously as sensors to acquire signals related to the health of patients, as processors to elaborate such signals and extract information, and as hubs to collect data from external medical devices and sensors. The processing techniques employed to obtain information about the health of patients but suitable to be implemented on board the smartphone are presented after that. In particular, audio, network interface, and accelerometer information processing is discussed. Conclusions are drawn in the final section.

Communication Architecture for Co-Morbidities Management

A communication architecture for co-morbidities management is aimed at allowing diverse medical devices such as sensors and actuators to interact within treatment scenarios tailored to the needs of co-morbid patients and also at improving the coordination of caregivers. The architecture should include location-independent interconnection, decision support, and partly automated functional adaptation according to the distinct needs of a patient. In practice, the communication architecture for co-morbidities management will be composed of different devices acting as one healthcare system to provide personalized care to patients at home, therefore improving social integration and quality of life. This solution will, at the same time, lead to lower costs. Medical decision-support and machine learning algorithms can be employed to orchestrate various components.

	Requirements						
	a)	b)	c)	d)	e)	f)	g)
Solutions							
[1]	Yes	Yes	No	Yes	No	No	No
[2]	Yes	Yes	No	Yes	No	No	No
[3]	Yes	Yes	Yes	Yes	No	No	No
[4]	Yes	Yes	No	No	No	No	No
[6]	Yes	Yes	No	No	No	No	No
[7]	Yes	No	No	No	No	No	No
[8]	Yes	Yes	No	Yes	No	No	No
[9]	Yes	Yes	No	Yes	No	No	Yes
[12]	Yes	Yes	No	Yes	No	No	Yes
This article	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 1. Comparison of some of the solutions in the literature.

A communication architecture to monitor health may be very useful in case of co-morbid patients who often are elderly and alone. In this view an efficient communication architecture for co-morbidities management should include the following requirements:

a) The presence of multi-sensors that monitor different health parameters that are essential both to check single pathologies and to have a general vision of the co-morbid patient health.
b) The capability to transmit the sensed parameter values remotely.
c) The possibility to set, modify, and control the action and the configuration of each single sensor remotely.
d) The possibility of medical and non-medical caregivers interacting with the patient remotely.

In addition, the architecture should include the following requirements linked to information processing capabilities:

e) To know if the patient is alone or not, possibly getting additional information about the environment where he/she is living, such as the number and identity of people at home, and the level of noise in the environment.
f) To identify the position of the patient both outdoors and indoors with a high degree of precision, such as a specific room within a house.
g) To recognize the type of physical activity the patient is performing, for example, walking, running, or sitting.

Table 1 shows a comparison of some of the solutions available in the literature. The architecture proposed in this article is the only one meeting all requirements, including the information processing-based requirements, i.e. e), f), and g).

This solution employs the smartphone simultaneously as a sensor to acquire signals related to the health of patients, as a processor to elaborate such signals and extract information of interest, and as a hub to collect other data from external medical devices.

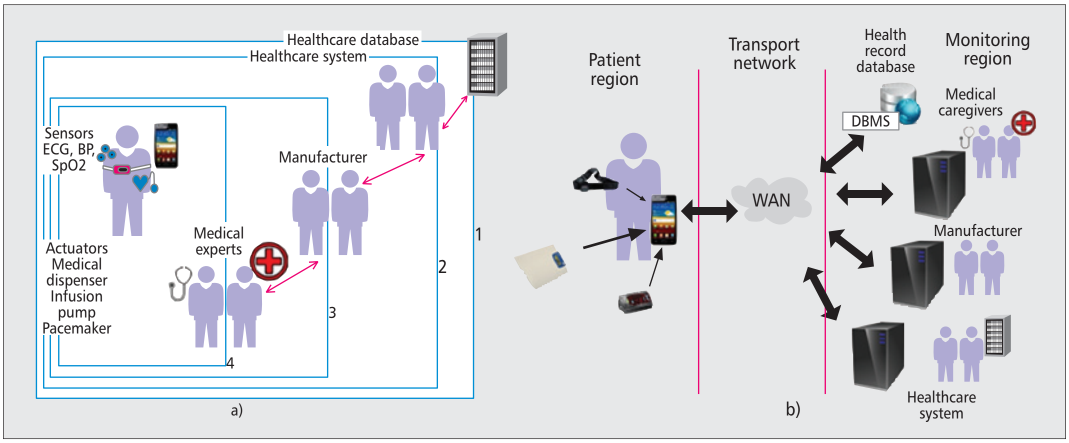


Figure 1. a) Integrated healthcare system; b) the smartphone-centric aal platform for co-morbidities monitoring.

The transfer of information may be structured into four groups. The presentation of relevant information should be adapted to the needs of the particular medical specialists and experts receiving information. Data communication groups can be represented by the loops in Fig. 1a. From the outer to the inner loop:

Device — Healthcare Database. Data detected by devices are forwarded to a remote database where they are memorized. Data available in the database might also be accessed from devices.

Device — Healthcare System. Information is utilized to control device functions. Time-relevant information such as remote warnings can be addressed to patients as well as to on-site healthcare personnel.

Device — Manufacturer. Data can be delivered in abstract and anonymous form to the device manufacturer. This information can be used for the development and refinement of next generation devices.

Device — Medical Experts. Medical experts access information via web-access services. They can affect the treatment system, change the therapy, and give medical advice displayed on user interfaces.

Moreover, medical experts, manufacturers, health system personnel, and the healthcare database may want to communicate each other (red arrows in Fig. 1a).

Smartphone-Centric Solution for e-Health Applications

A specific implementation of the general architecture presented in the previous section is represented by the smartphone-centric solution detailed in [9].

This solution employs the smartphone simultaneously as a sensor to acquire signals related to the health of patients, as a processor to elaborate such signals and extract information of interest, and as a hub to collect other data from external medical devices. In [9] this working modality is called the "hub+sensor+processor" paradigm. The smartphone transmits and receives such data by using several communication interfaces (e.g. WiFi, 3G/4G, GPRS). This smartphone-centric solution and its "hub+sensor+processor" action may be exploited to obtain a communication architecture for co-morbidities management with the features recommended in the previous section. The smartphone-centric choice allows reducing the number of required components and implementing ubiquitous, automatic, and precise monitoring. The block scheme composing the smartphone-centric architecture is shown in Fig. 1b.

In general, the architecture can be divided into three regions: patient, transport network, and monitoring. The patient region is usually a Personal Area Network (PAN) and may be composed both by wearable sensors that define a Body Area Network (BAN) as well as by non-wearable sensors. In the case shown in Fig. 1b, there are two non-wearable sensors: a pulse oximeter to measure the saturation of peripheral oxygen, and a scale to measure body weight. The smartphone is used by following the aforementioned "hub+sensor+processor" paradigm: it receives data from external sensors and manages data detected by embedded sensors, processes information as detailed in the next section, and conveys information through a Wide Area Network (WAN) to the final destination. In addition, smartphones can also send data to such external sensors, thus giving to the medical caregivers the possibility to set, modify, and control the action and the configuration of each single sensor remotely.

This approach exploits the great expansion of cellular communication networks and solves many problems concerning connectivity coverage. The authors in [6] outline three other critical factors for tele-monitoring platforms: usability, quality of transmitted data, and interference with other devices. The described smartphone-centric platform does not completely solve all such issues but represents a suitable solution.

The WAN is the transport network. It is a telecommunications network accessed through either a mobile phone network typically used for data by smartphones or through WiFi interfaces. The final destination, in the described platform, is the monitoring region and may be:

- A database server where the parameters of all monitored patients are stored and made available to medical staff, manufacturers and healthcare systems.
- Healthcare system personnel.
- Manufacturers.
- Medical experts, as discussed in the previous section.

Beyond the access to the database, the smartphone-centric architecture also assures seamless communication between the staff within the monitoring region.

Neither accelerometer nor localization external dedicated sensors are used because motion identification and localization are provided by using the smartphone itself. For example, as detailed in the next section, the accelerometer sensor embedded in the smartphone can be exploited to accomplish a precise recognition of the patient's physical activity. Currently adopted applications track distances and times covered during workouts and fitness activities by means of GPS receivers. The use of embedded accelerometers allows the detection of the type of performed physical activity because it can recognize specific movements.

Smartphones communicate with external devices in the patient region by using Bluetooth interfaces. In addition to the instruments shown in Fig. 1b, other useful sensor systems may be considered, such as: chest strips worn by patients, which can provide an approximation of the trans-thoracic impedance and of the heart rate; ElectroCardioGram (ECG); ElectroEncephaloGram (EEG); and glucose and blood pressure sensors. The smartphone-centric platform solves the open technical problem of interoperability, typical in tele-monitoring systems, between devices and the hub, and between the hub and the final destinations. Smartphones are already integrated with telecommunication network interfaces for data and voice transmission and, in many cases, already provide the necessary interfaces to connect to devices. In practice, using smartphones [4, 9, 10] simplifies the connection with sensors and the forwarding of measurements to the interworking network. Moreover, the presence of such a common device may facilitate the patient's acceptance of the monitoring system.

As stated, the viewpoint presented in this article is to give smartphones a new additional role: not only "hub+sensor" but "hub+sensor+processor." The idea of using a "hub+sensor" capability is in the literature (e.g. [4] and [10]). Sensor technologies combined with mobile communications were used to track patients' health measurements. Actually, sensors embedded into smartphones can be efficiently used for health monitoring: accelerometers can register different motions and walking gaits; infrared photo-detectors can measure body temperature; and more recently external sensors can be added to smartphones to measure heat flux, heart rate, and blood glucose levels. The additional "processing" capability may be an important added value for the platform. In summary, the platform offers:

- Both short-range (Bluetooth and WiFi employed for local information exchange) and long-range (GPRS, 3G/4G, and WiFi employed as Internet access) communication capabilities by exploiting the smartphone's network interfaces.
- Sensing capabilities: modern smartphones can acquire data by embedded sensors (always available on smartphones) such as GPS receivers, accelerometers, microphones, and radio interfaces often employed for localization purpose, or by external sensors such as pulse oximeters and scales connected through cables or local radio interfaces.
- Interaction capabilities, by exploiting the smartphone's multimedia features, useful to provide warnings, suggestions, and recommendations to patients when needed.
- Information processing capabilities provided by modern smartphones often equipped with different CPUs, efficient operating systems (such as Android), and flexible software. The following section reports on three important smartphone actions implemented over the proposed platform, audio recognition, localization, and physical activity detection, which are made feasible by the mentioned processing capabilities.

Information Processing Capabilities

We focus our description on smartphones' information processing capabilities that, from the authors' viewpoint, represent the new key function of the "hub+sensor+processor" paradigm and of the proposed AAL architecture. As stressed earlier, co-morbidity management implies not only multi-sensor acquisition and transfer, but also additional functions to know if the patient is alone or not and who is with him/her; to identify the location of the patient; and to recognize his/her physical activity.

Information about whether a person is alone or not, about the number of people with him/her, about their identity, and about the level of noise, may stem from an audio processing-based approach concerning speakers' count and recognition such as the one presented below.

Localization may derive from information processing based on signals received by smartphones' network interfaces. A possible algorithm is proposed below. In particular, the place recognition scheme proposed in [12] has been taken into account. Information required to carry out such a process is obtained from multiple sources such as the WiFi interface (in the case of indoor places) and the GPS receiver (in the case of outdoor places). The method, suitable for smartphone implementation, is briefly described and its recognition accuracy performance is presented for the specific case of monitoring a patient at home.

A physical activity recognition method based on raw data acquired directly from the measurements carried out by the smartphone accelerometer is reported below.

Audio Information Processing

In this context, audio processing over smartphones is aimed at classifying an audio signal acquired by embedded microphones so providing information about the noisy level of the

Information about whether a person is alone or not, about the number of people with him/her, about their identity, and about the level of noise, may stem from an audio processing-based approach concerning speakers' count and recognition such as the one presented below.

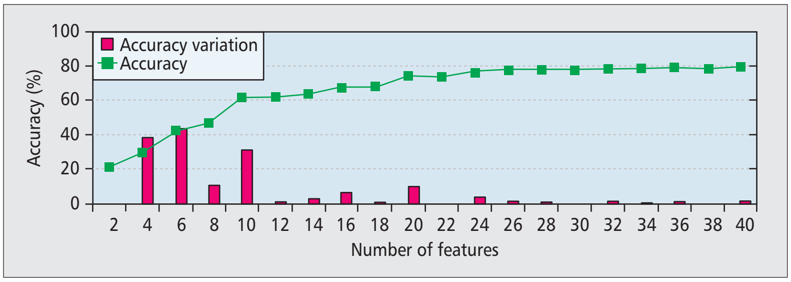


Figure 2. Accuracy of the speaker recognition method.

environment (e.g. excessive volume of audio/video equipment), and about the number and the gender of active speakers in the smartphone's surroundings. Audio processing may also be used to check the identity of a person, and this is the case briefly treated in the following. Actually, having information about the identity of caregivers will also provide information about the number of people with the patient and, implicitly, if he is alone or not, even if this information may also be retrieved through different algorithms in the literature. Several speaker recognition algorithms can be found in the literature, both for closed-set and open-set applications. Closed-set implies the classification of data belonging to a set of speakers known a priori, while in the open-set scenario there is no available knowledge on the set of speakers. In more technical detail, after the raw audio signal is acquired, the smartphone extracts a compact representation of the signal called *features*. Coherently with the state of the art in the field, the Mel Frequency Cepstrum Coefficients (MFCC) and their Delta Delta Coefficients (DDC) have been used in this article. The *features* feed a classifier that provides the result (i.e. the recognized speaker). Speaker recognition is implemented through a process divided into two phases, offline and online. The former aims at training the employed classifier, a Support Vector machine (SVM) in the case of this article, by using the *features* extracted from the speeches of different reference speakers (i.e. people expected to be in the patients surroundings), representing the predefined set. The latter is the recognition, where the *features* are extracted from an unknown speaker and used as input for the previously trained SVM. Finally, the unknown speaker is recognized as one of the speakers of the predefined set or he/she remains unidentified.

Figure 2 shows the accuracy, that is, the percentage of correct speaker recognitions of the described method versus the number of employed *features*. The aim of this test is recognizing a patient and five members of his/her family. In practice, six speakers have been used to train the SVM model. The green line represents the obtained accuracy. When the number of used *features* is above 26, the accuracy remains constant. Figure 2 also includes the accuracy variation histogram, which represents the accuracy gain obtained by increasing the number of *features* with respect to the previously tested value. For example, the employment of four *features* instead of two allows obtaining an accuracy increase close to 40 percent. Practically, the audio information processing capability of the smartphone-centric platform allows individuating the correct speaker within the set composed by the patient and his relatives in the 80 percent of cases.

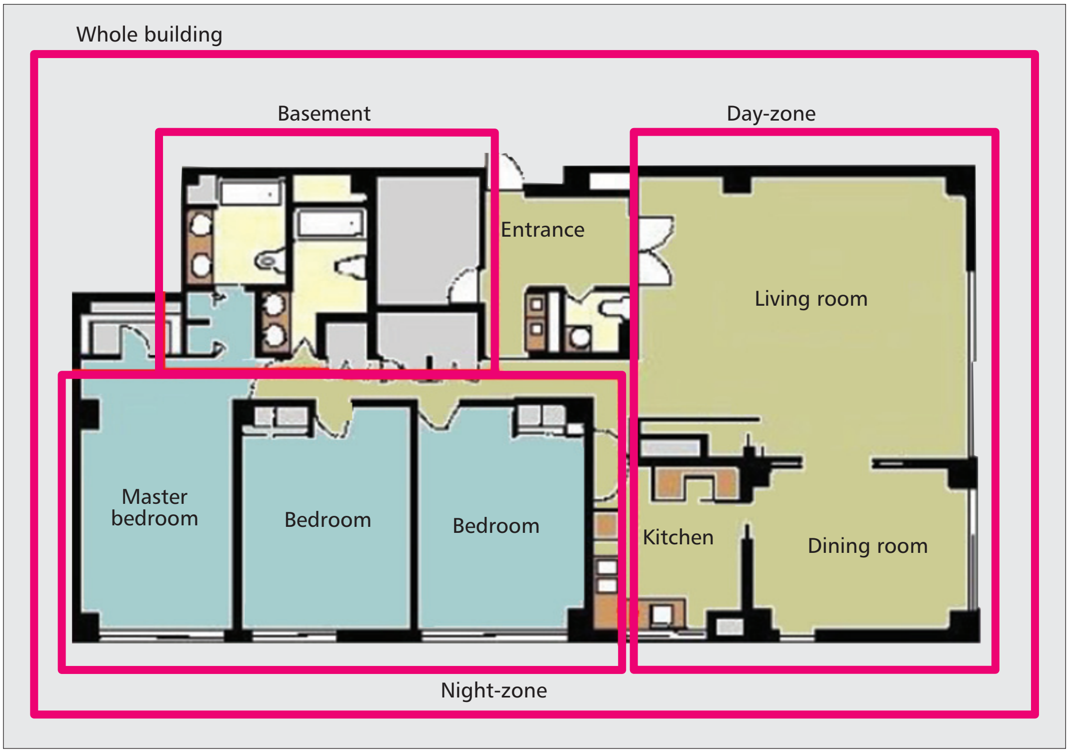


Figure 3. Considered localization places.

Network Interface Information Processing

Processing of signals received by smartphone network interfaces is the basis of Location-Based Services (LBSs), which are information services, accessible through mobile devices, such as Smartphones, that provide people and object localization. LBSs can be used in many applicative scenarios, such as health, object search, entertainment, work, and personal life. A well known localization process concerns the family of place recognition (PR) algorithms. The key idea of such algorithms is to recognize user localization not by identifying geographical coordinates but simply understanding in which place a user is staying (e.g. at home or at the gym). In this article we consider the Location Recognition Algorithm for Automatic Check-In algorithm (LRACI) [12] in the context of the AAL smartphone-centric platform. LRACI is employed to determine, in a completely transparent, automatic, and non-invasive way, in which room a patient is. The only request to the patient is having the smartphone with him/her. The performance of LRACI applied to a case in which there are four different places where the patient can be is reported in the following. The scenario is shown in Fig. 3. Localization output can be:

1. Day-zone (living and dining room, and kitchen) where a WiFi access point (AP) is installed.
2. Night-zone (bedroom).
3. Basement.
4. The whole building.
5. No location.

Whole building obviously contains places 1, 2, and 3, and is employed to recognize if the patient is at home but not within any of the first three places. No location represents the case in which the patient is not localized at home.

Also, this algorithm is based on offline and online phases. During the offline (or training) phase the patient's smartphone collects measurements related to GPS/HPS signals and/or to detected WiFi APs for each considered place. These measurements are then used in order to build the reference finger print (RFP) characterizing a place. RFPs are either stored remotely or directly on-board the smartphone. In the online (or recognition) phase, the smartphone collects the same measurements (GPS/HPS and/or WiFi AP signals) online and computes a finger print (FP) that is compared with the stored RFPs. The patient is localized in the place whose RFP is the closest to the acquired FP.

The obtained performance is reported in Fig. 4. It is the confusion matrix of the PR algorithm and shows the percentage of rooms correctly recognized during the tests. In the offline phase, the RFPs of the five considered locations have been built by collecting GPS/HPS and WiFi AP signals in 50 different points.

The percentage values reported in Fig. 4 have been computed by averaging the results obtained by 50 recognition phases. The performance is very satisfying. If the patient is in the day-zone, he/she is localized there in 94.5 percent of cases and confused with the night-zone in 5.5 percent of cases (first line in Fig. 4). The presence in the night-zone is correctly identified in 81.1 percent of cases and mistaken with the day-zone in 10.8 percent of cases and no location 8.1 percent of cases. The basement is recognized in 91.3 percent of cases and sometimes confused with no location (8.7 percent of cases). The presence within the whole building is correctly identified in 98 percent of cases and mistaken with no location in 2 percent of cases. No presence at home is recognized in 83.9 percent of cases and mistaken with the presence in the day-zone in 14.7 percent of cases and the night-zone in 2.3 percent of cases. The average accuracy is 89.8 percent. The accuracy is higher in the day-zone also thanks to the presence of an AP in the location. In general, when a location contains an AP, its radio signal dominates the others, characterizes the RFP of the location, and enables an efficient recognition. The absence of a dedicated AP causes a degradation of the location recognition accuracy. LRACI performance is not so satisfying when two adjacent locations must be discriminated. This is the case of the night-zone: about 11 percent recognitions are not correct because the night-zone is confused with the adjacent day-zone. This problem happens when WiFi signals are shared. The accuracy obtained for whole building is high because, in this specific case also, GPS/HPS positioning information can be efficiently used.

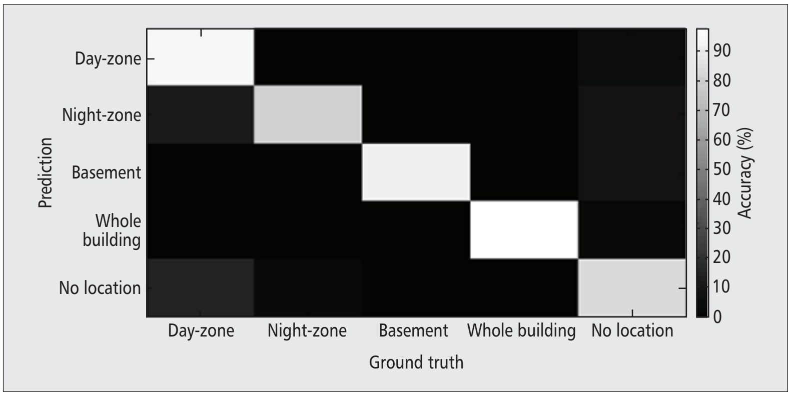


Figure 4. Accuracy of the place recognition represented by the confusion matrix.

Accelerometer Information Processing

The last information processing capability considered in this article concerns the physical activity recognition of the co-morbid patients. It is based on the action of sensing, processing, and classification of the signal provided by the smartphone-embedded accelerometer. The algorithm is designed to recognize eight different classes of physical activities: idle, sitting, standing, walking, going up and down the stairs (contracted in upstairs and downstairs), running, and cycling. These classes are particularly useful in case of cardio circulatory pathologies.

Again, two phases, offline/training and online, are the basis of the processing procedure. The acquisition of training signals is performed by keeping the smartphone in different positions. The algorithm periodically collects the raw signal from the smartphone accelerometer and organizes it into frames. A feature vector is computed for every frame and is used by a classifier, in this case

Features set		Kilometers	
Activities	**Percentage (%)**	**Activities**	**Percentage (%)**
Cycling	84.9	Cycling	60.3
Downstairs	69.5	Downstairs	0
Idle	79.2	Idle	0
Running	99.2	Running	96.1
Sitting	98.4	Sitting	85.5
Standing	91.2	Standing	47.3
Upstairs	63.1	Upstairs	70.7
Walking	51.1	Walking	54.7
Average	**79.5**	**Average**	**51.8**

Table 2. Accuracy of the activity recognition algorithm.

a decision tree (DT), to classify the frame into one of the movement classes previously listed.

In order to determine the best classification accuracy of the movements, numerous features were evaluated and compared: mean, zero crossing rate, energy, standard deviation, cross-correlation, sum of absolute values, sum of variances, and number of peaks of the signal obtained from the accelerometer. The *feature* vector chosen for the tests shown in this article is made of nine features (i.e. mean, standard deviation, and number of peaks of the accelerometer measurements along the three axes) as in [9]. This approach is identified as "*Features Set*." The obtained results have been compared with the approach in which only one *feature* has been used: the Km parameter, strictly related to the energy of the accelerometer signal, proposed and detailed in [10]. Such comparison is proposed since the solution reported in [10] is one of the reference architectures applicable to co-morbid patients monitoring scenarios.

The training signals employed in the tests have been acquired by four volunteers. Each volunteer acquired approximately one hour signal for each of the classes listed above. In order to determine the performance, the accelerometer signal has been acquired by a fifth volunteer not involved in the training phase.

Table 2 shows that "*Features Set*" allows obtaining a more accurate and precise classification of the patient's movements with respect to "*Km*." The average accuracy is approximately 80 percent if the "*Features Set*" is employed, while it is approximately 52 percent if "*Km*" is used.

Conclusions

This article describes the main characteristics of a smartphone-centric Ambient Assisted Living (AAL) platform aimed at monitoring, at home, patients suffering from a set of physical and mental limitations, called co-morbidities. The article highlights that smartphones have both short-range and long-range communication capabilities; information processing capabilities; and sensing capabilities implemented by internal and external sensors. The specific case of co-morbidities management implies the following needs: to acquire data from a set of sensors that monitor different health parameters; to transmit the acquired values remotely; and to control the action and configuration of single sensors. Moreover, an efficient communication architecture for co-morbidity monitoring and management should also assure the possibility of healthcare staff to interact with each other and with the patient remotely, and should guarantee the power to know if the patient is alone or not and who are the caregivers, to localize the patient, and to identify the physical activity performed by the patient. As a consequence, this article is focused on the information processing capabilities of the smartphones with particular emphasis on audio, network interfaces, and accelerometer information processing. These kinds of information can help monitor co-morbid patients remotely. The presented solutions have been designed and practically implemented by using off-the-shelf smartphones. In more detail, the following solutions have been presented: an audio processing-based approach, aimed at recognizing the identity of people who are with the monitored patients at a given time, which implicitly helps to monitor if a patient is alone; a place recognition method where the required information is obtained from multiple sources such as the smartphone WiFi interface, in the case of indoor localization, and the GPS receiver, in the case of outdoor localization; and a physical activity recognition method based on raw data directly acquired from the smartphone accelerometer. In all cases a brief presentation of the performance has been provided. The obtained results allow concluding that the employed information processing solutions are reliable and suitable to be employed in the described AAL smartphone-centric platform for co-morbidity monitoring.

References

[1] S. Adibi, "Link Technologies and BlackBerry Mobile Health (mHealth) Solutions: A Review," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no 4, Jul. 2012.

[2] C. C. Y. Poon, Y.-T. Zhang, and S.-D. Bao, "A Novel Biometrics Method to Secure Wireless Body Area Sensor Networks for Telemedicine and M-Health," *IEEE Commun. Mag.*, April 2006.

[3] A. J. Jara, M. A. Zamora-Izquierdo, and A. F. Skarmeta, "Interconnection Framework for mHealth and Remote Monitoring Based on the Internet of Things," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013.

[4] R. Carroll *et al.*, "Continua: An Interoperable Personal Healthcare Ecosystem," *IEEE Pervasive Computing*, vol. 6, no. 4, Oct.-Dec. 2007.

[5] M. Chen *et al.*, "Body Area Networks: A Survey," *ACM/Springer Mobile Networks and Applications* (MONET), Feb. 2011, DOI: 10.1007/s11036-010-0260-8.

[6] L. Pecchia, P. Melillo, and M. Bracale, "Remote Health Monitoring of Heart Failure with Data Mining via CART Method on HRV Features," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, March 2011, pp.800–04, doi: 10.1109/TBME.2010.2092776.

[7] O. R. E. Pereira, J. M. P. L. Caldeira, and J. J. P. C. Rodrigues, "Body Sensor Network Mobile Solutions for Biofeedback Monitoring," *Mobile Networks and Applications* (MONET), Springer, ISSN: 1383-469X (print), ISSN: 1572-8153 (electronic), vol. 16, no. 6, Dec. 2011, pp. 713-732, doi: 10.1007/s11036-010-0278-y.

[8] E. Villalba *et al.*, "Wearable and Mobile System to Manage Remotely Heart Failure," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 6, Nov. 2009, pp. 990–96.

[9] I. Bisio *et al.*, "A Smartphone-Centric Platform for Remote Health Monitoring of Heart Failure," *Wiley International Journal of Communication Systems*, Article first published online: 14 APR 2014, DOI: 10.1002/dac.2778.

[10] M. K. Suh *et al.*, "A Remote Patient Monitoring System for Congestive Heart Failure," *J. Medical Systems* (JOMS), May 2011.

[11] C.-L. Chen, C.-C. Lee, and C.-Y. Hsu, "Mobile Device Integration of a Fingerprint Biometric Remote Authentication Scheme," *Int'l J. Commun. Systems*; online 28 April 2011 in Wiley Online Library, doi: 10.1002/dac.1277.

[12] I. Bisio *et al.*, "GPS/HPS- and WiFi Fingerprint-Based Location Recognition for Check-In Applications over Smartphones in Cloud-based LBSs," *IEEE Trans. Multimedi*a, vol. 15, no. 4, June 2013, pp. 858–69, doi: 10.1109/TMM.2013.2239631.

Biographies

IGOR BISIO (igor.bisio@unige.it) was born in Novi Ligure, Italy in 1978. He received his "Laurea" degree in telecommunication engineering at the University of Genoa, Italy in 2002 and his Ph.D. in 2006. He is currently an assistant professor and member of the Digital Signal Processing (DSP) and Satellite Communications and Networking (SCNL) laboratories in the DITEN Department at the University of Genoa. He is the author of approximately 100 scientific papers, including international journals, international conferences, and book chapters. His main research activities concern signal processing over portable devices such as smartphones, context and location awareness, adaptive coding mechanisms, indoor localization, security and e-health applications, resource allocation and management for satellite and space communication systems.

MARIO MARCHESE (mario.marchese@unige.it) was born in Genoa, Italy, in 1967. He received the "Laurea" degree (cum laude) in electronic engineering and the Ph.D. in telecommunications from the University of Genoa, Genoa, Italy, in 1992 and 1996, respectively. He is currently an associate professor with the DITEN Department, University of Genoa. He is the founder of and responsible for the Satellite Communications and Networking (SCNL) Laboratory at the DITEN Department of the University of Genoa. He is the author of approximately 250 scientific papers including international journals, international conferences, book chapters, and the book *QoS over Heterogeneous Networks*. His main research activity concerns satellite and radio networks, transport layer over satellite and wireless networks, quality of service and data transport over heterogeneous networks, and applications for smartphones.

FABIO LAVAGETTO (fabio.lavagetto@unige.it) was born in Genoa in 1962. He is currently a full professor in telecommunications at the DITEN Department of the University of Genoa. He was vice-chancellor with responsibility for research and technology transfer at the University of Genoa. Since 2005 he has been vice-chair of the Institute for Advanced Studies in Information Technology and Communication. Since 1995 he was the head of research of the Digital Signal Processing (DSP) Laboratory of the University of Genoa. He was general chair of several international scientific conferences and has authored over 100 scientific publications in international journals and conferences. His main research activities concern signal processing over portable devices such as smartphones, context and location awareness, adaptive coding mechanisms, indoor localization, security and e-health applications.

ANDREA SCIARRONE (andrea.sciarrone@unige.it) was born in Livorno, Italy in 1984. He received his bachelor degree in telecommunication engineering at the University of Genoa in 2007; in 2009 he achieved a master degree cum laude in telecommunication engineering at the same university. In April 2014 he received is Ph.D. in science and space engineering with a thesis on the processing of heterogeneous signals for context-aware applications. Currently he is a research fellow at the DITEN Department of the University of Genoa. His main research activity concerns signal processing over portable devices such as smartphones, context and location awareness, indoor localization, security, e-health applications and the android operating system.

Overcoming the Challenge of Variety: Big Data Abstraction, the Next Evolution of Data Management for AAL Communication Systems

Rui Mao, Honglong Xu, Wenbo Wu, Jianqiang Li, Yan Li, and Minhua Lu

ABSTRACT

With the extensive use of information technology in AAL communication systems, a data management model has recently embodied the 3-V characteristics of big data: volume, velocity, and variety. A lot of work has been done on volume and velocity, but not as much has been reported on variety. To handle the variety of data, universal solutions with acceptable performance are usually much more cost effective than customized solutions. To achieve universality, a basic idea is to first define a universal abstraction that covers a wide range of data types, and then build a universal system for universal abstraction. Traditional database management systems commonly use a multidimensional data type, or feature vectors, as a universal abstraction. However, many new data types in AAL systems cannot be abstracted into multidimensional space. To find a more universal data abstraction and build more universal systems, we propose the concept of big data abstraction, with metric space as a universal abstraction for AAL data types. Furthermore, to demonstrate how metric-space data abstraction works, we survey the state of the art in metric space indexing, a fundamental task in data management. Finally, open research issues are discussed.

INTRODUCTION

Today, with the extensive use of information technology, data management and analysis for ambient assisted living (AAL) communication systems has gradually come to embody the 3-V characteristics of big data [1]:

- Volume: The amount of data in regard to computation and storage is extremely large.
- Velocity: The speed of data input and output is extremely high.
- Variety: The range of data types and sources is extremely wide.

As a result, AAL data has grown beyond the capability of most available database management tools or traditional data processing applications. A revolutionary approach to big data management in AAL is in great need.

Scholars and practitioners all over the world have done considerable intensive research on big data. However, most of the effort was spent on volume and velocity, but not as much on variety.

Some common data types in AAL communication systems are listed in Table 1.

To cope with the large number of data types, there are two basic types of solutions, customized ones and universal ones. Customized solutions build a customized system for each individual type of data, while universal solutions build a single system that can support a wide range of data types. If the performance is acceptable to the applications, universal solutions are much more cost effective. As a result, most commercial database management systems (DBMSs) are universal solutions so that they can be sold to many customers to maximize the profit.

A basic question is how to develop universal solutions. Looking back on the history of DBMSs, one can discover the basic paradigms of universal solutions. That is, one first defines a universal abstraction that covers a wide range of data types, and then builds a universal system for the universal abstraction based on its properties. Since every particular data type is a special case of the universal abstraction, a solution to the universal solution works for any data type it covers.

Commonly, traditional DBMSs make use of a multidimensional data type as a universal abstraction. That is, most data types are essentially represented by one or multiple numbers (i.e., a feature vector). However, many new data types in AAL systems cannot be abstracted into multidimensional space, and a more universal abstraction is needed for AAL data.

We propose a new concept of big data abstraction using metric space as the universal abstraction for AAL data types. Informally, a metric space [2] is a set with a distance function defined on its elements, where the distance function satisfies the triangle inequality. We show that metric space is more universal than multidi-

Rui. Mao, Honglong Xu, Yan Li, and Minhua Lu (corresponding author) are with Shenzhen University.

Wenbo Wu is with the University of Georgia.

Jianqiang Li is with Beijing University of Technology.

0163-6804/15/$25.00 © 2015 IEEE

mensional space and covers a wide range of AAL data types.

Big data abstraction is in its early stage of development. To demonstrate how big data abstraction might work, we survey the state of the art in metric space indexing. Indexing, or searching, is one of the fundamental tasks of data management and analysis. A lot of work has been done on metric space indexing. We believe what has been done for indexing provides excellent hints for other data management and analysis tasks. Open issues of big data abstraction in theory and application are also discussed.

The rest of this article is organized as follows. A discussion of customized and universal solutions is presented in the following section. After that we propose big data abstraction, and survey the state of the art in metric space indexing. In the final section, open research issues are discussed.

Universalization: Why and How?

In this section, we first show the necessity of universalization by comparing customized and universal solutions, then show the basic approach to achieving universality by reviewing the history of data management systems, and last discuss the current status of big data management with respect to variety.

Universalization: Why?

Facing various data types, customized solutions build one system for each data type. Since the system is tailored for a single data type, its performance can be expected to be high. However, its range of applicability is relatively narrow, and its price is thus relatively high. As a result, the performance-price ratio will be relatively low, and less profit has to be expected.

Universal solutions, on the contrary, build one system to support a wide range of data types. After fine tuning, the performance of universal systems is generally acceptable, except for some performance-critical applications. Because of its wide applicability, a universal system can be sold to many customers at relatively low prices. As a result, universal systems are more cost effective and more profitable.

Customized solutions are more suitable for performance-critical applications, while universal solutions achieve better balance between performance and price. Usually, buyers of AAL data management systems prefer universal solutions because of their low prices, given that the performance is acceptable. Likewise, providers of AAL data management systems tend to develop universal solutions to gain more customers and profit. Consequently, universal solutions are more popular than customized solutions in practice. The relationship between customized and universal solutions is similar to that between tailor-made and factory-made clothes. The next question is how to achieve universality.

Universalization: How?

Let us look back on the history of data management systems (Fig. 1), which always show an evolutionary trend from customization to universalization.

Data category	Data type
Behavioral habit data	Sleep time, frequency of wake up, restroom time and frequency, shower time, eating time, walking speed, time in and out
Physiological information	Blood pressure, blood lipids, blood oxygen, temperature, pulse, BMI, weight, bone density, respiratory rate
Healthcare information	Gene sequence, protein sequence, medical image
Environmental data	Surveillance video, noise level, pollution density, weather conditions

Table 1. Common data types in AAL communication systems.

In the early 1960s (Fig. 1a), with the increasing application of computers in enterprise management, large businesses began to build their own enterprise information systems, where common data were numbers: employee IDs, product prices, and so on. Since these systems were only used inside businesses, many of them were built, and a lot of resources were consumed. In the 1970s, the B-tree index was designed and integrated into relational DBMSs. B-tree supports search of numeric values, whether natural numbers, integers, or real numbers. That is, 1D data served as an abstraction of natural numbers, integers, or real numbers, and B-tree worked for all these data types since they are all special cases of 1D data. Furthermore, the integration of SQL made relational DBMSs even easier to use. Thus, some businesses were attracted to relational DBMSs. As the number of customers increased, the price of relational DBMSs dropped. Consequently, businesses gradually replaced their own information systems with relational DBMSs for acceptable performance at a much lower price. This was the first evolution of data management systems from customization to universalization (Fig. 1a).

Figure 1b shows the second stage of evolution when manmade satellites were launched. To manage spatial information acquired by satellite, individual spatial data management systems were built. Spatial data are usually represented by feature vectors and matched by similarity defined by distance functions. Again, a lot of efforts were spent on building individual systems. Later, in the 1980s, multidimensional indexing such as R-tree and kD-tree were designed and integrated into relational databases. Multidimensional indexing supports similarity search of multidimensional data with Euclidean distance or alike. Furthermore, SQL was also extended to support spatial data type and similarity query. As a result, individual spatial data management systems were gradually replaced by spatial DBMSs. This was the second evolution of tje data management system from customization to universalization (Fig. 1b).

Studying the above two evolutions, one can summarize the basic approach to achieve universality into three steps:

1. Find a universal data types that cover various data types.
2. Find a universal distance function that covers various distance functions.

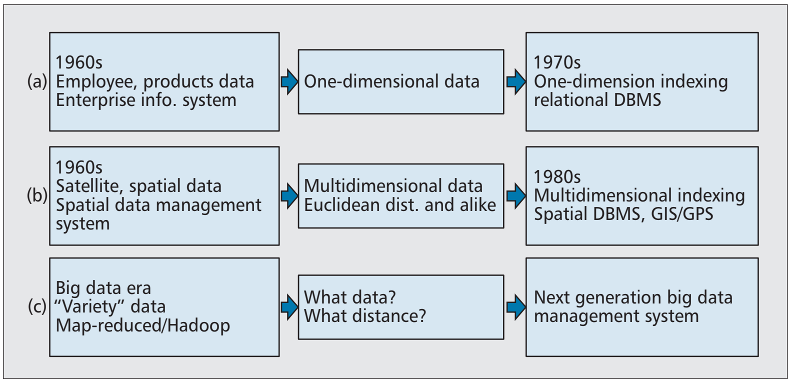


Figure 1. History of data management systems.

3. Build a system based on the properties of the universal data type and the universal distance function.

Discussion

In the second evolution, the multidimensional data type serves as the abstraction for spatial data types. A data type must satisfy two conditions to be covered by this abstraction:

- Data must be in feature vector form.
- Similarity of data must be defined by Euclidean distance or the like.

However, in the big data era, many data types and corresponding distance functions do not satisfy the above two conditions, such as text with edit distance, protein sequence with global alignment, or MMR images with Hausdorff distance. A new abstraction that can cover more AAL data types is in great need.

Map-Reduce is a very popular programming model to tackle big data applications these days. However, one should keep in mind at least two issues about map-reduce.

Who Developed It? — Map-Reduce was originally developed by Google, who possesses a great amount of both intelligence and resource to develop it. Building a big data system under the umbrella of the Map-Reduce model requires great programming skills and huge resource investment. Therefore, it is not for end users.

What Is It For? — Map-Reduce was originally developed to scan logs. After years of development, its functionality is still very limited. It is not an accurate claim that Map-Reduce has outperformed traditional DBMSs, but it is better only in limited application environments.

Therefore, we can conclude that the current status of AAL big data management is very similar to the early stage of building individual systems of the former two evolutions, and a new abstraction for big data is necessary to carry on the third evolution of data management systems from customization to universalization.

Big Data Abstraction

We propose the use of metric space as a universal abstraction for AAL data types, and to build a universal big data management and analysis system based only on the properties of metric space.

Metric Space

Informally, metric space is a set with a distance function, satisfying the triangle inequality, defined by its elements.

Definition — A metric space [2] is a pair (S, d), where S is a nonempty set and d is a real-valued distance function with the following properties:

For all $x, y \in S$, $d(x, y) \geq 0$ and $d(x, y) = 0$ iff $x = y$. (Positivity)

For all $x, y \in S$, $d(x, y) = d(y, x)$. (Symmetry)

For all $x, y, z \in S$, $d(x, y) + d(y, z) \geq d(x, z)$. (Triangle inequality)

Metric space requires only a metric distance function. An interpretation of the data in a coordinate system is not necessary. Two immediate advantages of using metric space as a universal abstraction are:

- Metric space is more universal than multidimensional space. Since Euclidean distance satisfies positivity, symmetry, and triangle inequality, multidimensional data with Euclidean distance form a special case of metric space. Some data types that cannot be abstracted into multidimensional space can be abstracted into metric space.
- A universal programming model can be built on metric space. To perform big data management and analysis, users only need to define their own data type and associated metric distance function, which can be plugged into the universal model as a black box.

Metric Space's Range of Application

Common AAL data types that can be abstracted into metric space are listed in Table 2. Except for examples 1 and 2, the examples cannot be directly abstracted into multidimensional space. Until now, only customized solutions have been developed for them.

For data types that cannot be directly abstracted into metric space, there are some alleviations. First, there are some mathematical approaches to convert non-metric distance functions to metric ones. Second, there are universal approaches for distance functions satisfying only some of the metric properties (e.g., semi-metric and pseudo-metric). Third, as long as some kind of inference can come from the distance function, universal approaches can be developed. An example is the protein identification problem with mass spectra. For three data objects x, y and z, an upper bound of $d(x, y)$ can be determined given $d(x, z)$ and $d(y, z)$, and a metric space index was adapted to support similarity queries of mass spectra [3].

The great universality of metric space is also one of the disadvantages of metric space abstraction. Domain-specific information is discarded. The triangle inequality of the distance function is the only property that can be leveraged. The key point is to recognize the pattern encapsulated by the distance function.

Metric-space-based big data abstraction is in its early stage of development. Among the fundamental tasks of metric space big data management and analysis, search is the only one that has received intensive research. To show the basic idea of how data management and analysis can be done in metric space, in the next section, we survey the state of the art in metric space indexing that supports similarity queries.

	Data type	Distance function
1	Number (one-dimensional)	Absolution value of difference
2	Vector (multidimensional)	Euclidean distance or alike
3	Text	Edit distance
4	Protein sequence	Global alignment (weighted edit distance)
5	Image	Hausdorff distance
6	Video	Percentage of similar frames

Table 2. Examples of metric space.

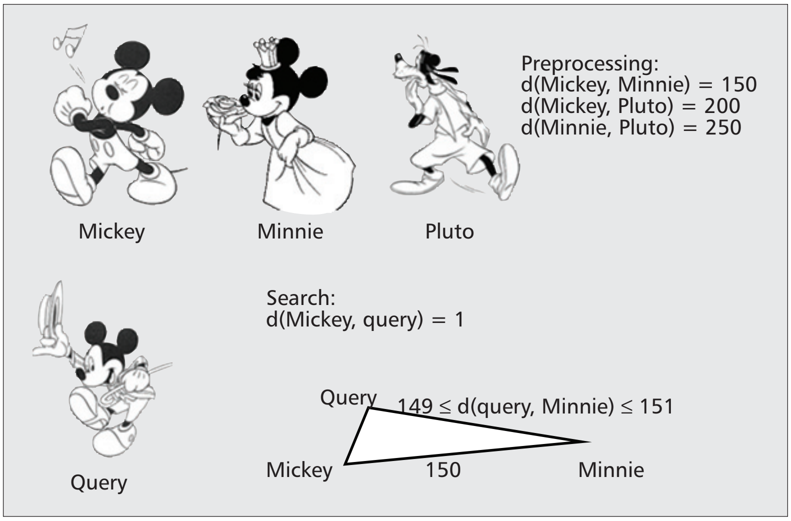


Figure 2. Triangle inequality: the principle of metric space abstraction.

Metric Space Indexing: A Universal Indexing for Similarity Queries

Given a database of data objects, a distance function as the similarity measurement, and a query object, a similarity query finds all data objects that are similar, determined by the distance function, to the query object.

Figure 2 illustrates the basic idea of how triangle inequality can be leveraged to answer similarity queries in metric space. Assume an image database consists of three cartoons of Mickey, Minnie, and Pluto, respectively [4]. Another cartoon of Mickey is used as a query [4], and we want to find all similar cartoons to it in the database. Since the distance calculation is usually costly for complex data types such as image, one goal is to minimize the number of distance calculations during the search. During preprocessing, pair-wise distances of the three cartoons in the database are calculated and stored. When the query comes, d(Mickey, query) is first calculated. Since the query is also a cartoon of Mickey, we can assume that d(Mickey, query) is small, say 1. Then, from the triangle inequality, it can be derived that 149 ≤ d(query, Minnie) ≤ 151 and 199 ≤ d(query, Pluto) ≤ 201. Therefore, neither Minnie nor Pluto is a query result. In a word, using triangle inequality, the similarity query is answered with only one distance calculation.

In the following, we first introduce the concept of an index, then survey common tree structured metric space indices and discuss their problems. Next, the pivot space model, a theoretical framework for metric space indexing, is introduced.

Index

A database index, or simply index, is a data structure to improve the efficiency of data lookup in a database. Answering similarity queries usually consists of two steps.

Offline Construction — Given the data set, construction builds an index data structure offline. The tree structure is one of the most popular metric space indexing structures. In their top-down construction, tree structure metric space indexing methods build index trees by recursively applying two basic steps: pivot selection and data partitioning. In pivot selection, a small number of reference points, called pivots, are selected from the database. In data partitioning, data points are partitioned by their distances to the pivots.

Online Search — Based on the offline built index data structure, similarity queries are answered online. The search process basically descends the index tree from the root to the leaf nodes. At each internal index node, some computation is performed based on the query and the information previously stored in the node. Some children are determined to be unable to contain any query results and can be pruned. Children that cannot be pruned are further visited in the same way. At each leaf index node, similarly, computation is performed to decide which data objects can be pruned and which data objects are query results without calculating their distances to the query. For the remaining data objects, their distances to the query are calculated to find the query results.

Common Metric Space Indexes

Based on the way data is partitioned, there are two kinds of metric space indices: the vantage point tree (VPT) and general hyper-plane tree (GHT) [5, 6].

During the offline construction of a VPT, the data space is partitioned into disjoint regions recursively. In each recursion, a vantage point, or pivot, is first selected as a reference point. Then the distances from all remaining points to the vantage point are calculated, and the median, m, of the distances is determined. Next, data is partitioned such that points with distances smaller than or equal to m move to the left

Many traditional metric space indexes only show their superiority to others experimentally. A theoretical framework is needed to analyze and compare different indexes and to predict the performance of new indexes.

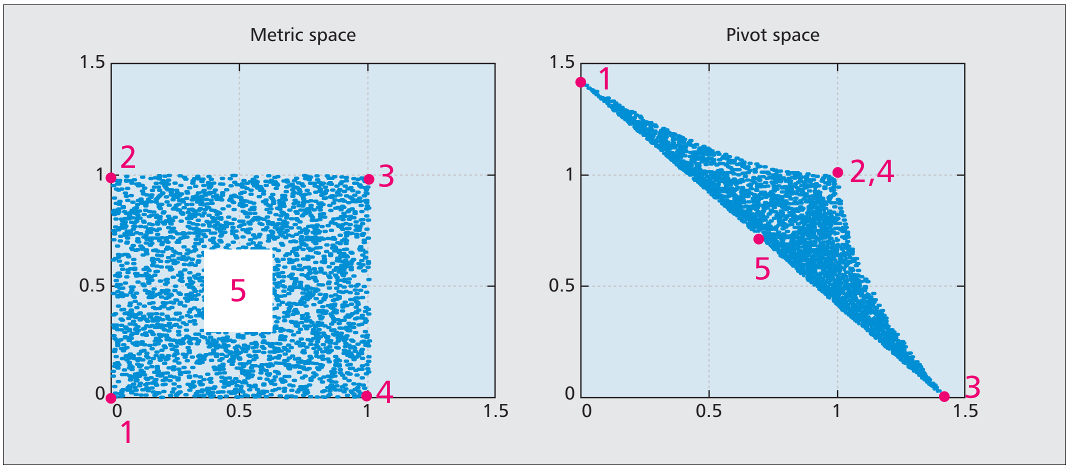


Figure 3. The pivot space of points from unit square with Euclidean distance.

child, and points with distance larger than m move to the right child.

In each recursive construction step of GHT, two points, c_1 and c_2, are selected as centers, or pivots. Each remaining point is assigned closer to the center. That is, points closer to c_1 than to c_2 are assigned to the left child, while points closer to c_2 than to c_1 are assigned to the right child.

Since metric space has no coordinate system, mathematical tools of multidimensional space are not directly applicable to metric space. Consequently, most of above work is based on heuristics. Theoretical analysis is usually overlooked. Nevertheless, a large amount of traditional metric space indices only show their superiority to others experimentally. A theoretical framework is needed to analyze and compare different indices and to predict the performance of new indices. The pivot space model introduced next aims to solve the above problems.

Pivot Space Model: A Theoretical Framework for Metric-Space Indexing

A first goal of the pivot space model is to impose coordinates to data in metric space so that mathematical tools of multidimensional space can be applied for theoretical analysis. To do so, a mapping from metric space to multidimensional space is defined [7].

Let R^m denote a general real coordinate space of dimension m. Let (S, d) be a metric space where $S = \{x_i \mid i = 1, 2, \ldots, n\}$ is the database, and d is a metric distance function. Let $P = \{p_j \mid j = 1, 2, \ldots, k\}$ be a set of k pivots. $P \subseteq S$. Duplicates are not allowed.

Given the set of pivots, each point x in S can be mapped to a point x_p in the non-negative orthant of R^k. The jth coordinate of x_p represents the distance from x to p_j:

$$x_p = (d(x, p_1), \ldots, d(x, p_k)).$$

The **pivot space** [7] of S, $F_{P,d}(S)$, is defined as the image of S:

$$F_{P,d}(S) = \{x_p \mid x_p = (d(x, p_1), \ldots, d(x, p_k)), x \in S\}.$$

Figure 3 gives an example of pivot space. The original data consists of 3000 points uniformly distributed in the unit square. With Euclidean distance, these points form a metric space. Two points, with coordinates (0,0) and (1,1), are selected as the pivots. Since the number of pivots is 2, the pivot space is also 2D. For a point, (a, b), in the original metric space, its coordinate on the x-axis in the pivot space is its Euclidean distance to the first pivot, (0, 0), and its coordinate on the y-axis in the pivot space is its Euclidean distance to the second pivot, (0, 1). Four special points (corners and the center) in the original metric space are selected, and their images in the pivot space are marked.

A **complete pivot space** [7] is a pivot space with all points selected as pivots. It has been proved that the mapping from metric space to a complete pivot space is isometric [7]. That is, the mapping is one-to-one, and the metric space distance of any pair of points equals the L_∞ distance of their images in the complete pivot space. Therefore, instead of the original metric space, the complete pivot space can be searched to answer similarity queries. Since the complete pivot space is a multidimensional space, the problem turns into a multidimensional indexing problem, to which many mathematical tools can be applied.

Furthermore, it has been proved that the partition boundaries of both VPT and GHT are straight lines (or hyper-planes) in the pivot space, with different slopes. In other words, VPT partition and GHT partition are essentially rotations of each other [8]. Moreover, it has been shown theoretically and experimentally that VPT outperforms GHT [8].

Open Issues and Future Work

Based on the survey of metric space indexing, one can deduce a possible paradigm for metric-space-based big data management and analysis tasks other than indexing. That is, we can first map data from metric space to pivot space, and then apply traditional multidimensional space mathematical tools to the pivot space.

Open research issues and future work include, but are not limited to, the following.

Given the isometric mapping, can complete pivot space replace metric space? The pivot space model defines an isometric mapping from metric space to the complete pivot space. Under this mapping, the pair-wise distances are preserved. For indexing, the complete pivot space can be searched instead of the original metric space. For other tasks (e.g., clustering and classification), can metric space be replaced by the complete pivot space? More theoretical analysis is to be performed on this topic.

If the above is "yes," can we work on the complete pivot space directly? Since all points are selected as pivots, the dimension of the complete pivot space equals the size of data, which is usually huge for big data applications. High dimensionality causes problems for many data management and analysis tasks, such as indexing. It is necessary to identify which tasks can be performed on the complete pivot space directly, and which cannot.

If the above is "no," how should dimension reduction for the complete pivot space be done? According the pivot space model, dimension reduction for the complete pivot space can only select existing dimensions and cannot create new dimensions for indexing. More theoretical analysis is to be done to determine for which data management and analysis tasks that dimension reduction for the complete pivot space model can create new dimensions.

How can performance be improved algorithmically? It is difficult to build a universal system generally performing well. Since metric space abstraction discards domain-specific information and only leverages the triangle inequality of the distance function, it is of key importance to refine the algorithms to achieve acceptable performance.

How can it be done in a parallel or distributed manner? Another way to improve performance is to exploit multiple processors. Parallel and distributed techniques are of great value.

How can metric distance functions be defined for more data types and applications? As discussed earlier, not all data types and distance functions can be abstracted into metric space. Defining proper metric distance functions for data types and applications is of critical importance.

CONCLUSIONS

This article focuses on the variety challenge of big data problems in AAL communications systems. First we show that a universal approach is very effective in overcoming variety. Then we show that universality can be achieved by abstraction. Next, metric space is proposed as a universal abstraction for AAL big data. To demonstrate how a metric space data management and analysis system can be built, we survey the state of the art in metric space indexing and introduce the pivot space model. Last but not least, a few important open research issues, which form the direction of future work, are discussed.

Since much less attention has been paid to variety than to volume and velocity, this article provides a novel perspective on designing AAL big data management and analysis systems.

ACKNOWLEDGMENT

This research was supported by the following grants: China 863: 2012AA01A309; NSF-China: 61170076, U1301252, 61471243; Shenzhen Foundational Research Projects SGLH20131010163759789, JCYJ20130401111833183, JCYJ20140418095735561.

Based on the survey of metric-space indexing, one can deduce a possible paradigm for metric space based big data management and analysis tasks other than indexing. That is, we can first map data from metric space to pivot space, and then apply traditional multi-dimensional space mathematical tools to the pivot space.

REFERENCES

[1] D. Laney, "3D Data Management: Controlling Data Volume, Velocity and Variety," Gartner, Feb. 2001.

[2] J. Matousek, *Lectures on Discrete Geometry*, Springer-Verlag, 497, 2002.

[3] S. R. Ramakrishnan *et al.*, "A Fast Coarse Filtering Method for Protein Identification by Mass Spectrometry," *Bioinformatics*, vol. 22, no. 12, 2006, pp. 1524–31.

[4] http://www.tom61.com/shaoertuku/jianbihuatupian/2010-12-18/811.html, Oct. 2014.

[5] E. Chavez, *et al.*, "Searching in Metric Spaces," *ACM Computing Surveys*, vol. 33, no. 3, 2001, pp. 273–321.

[6] P. Zezula *et al.*, *Similarity Search: The Metric Space Approach*, Springer, 2006.

[7] R. Mao, W. Miranker, and D. P. Miranker, "Pivot Selection: Dimension Reduction for Distance-Based Indexing," *J. Discrete Algorithms*, Elsevier, 2012, pp. 32–46.

[8] R. Mao *et al.*, "On Data Partitioning in Tree Structure Metric-Space Indexes," *Proc. 19th Int'l Conf. Database Systems for Advanced Applications*, Apr. 21–24, 2014, Bali, Indonesia, pp. 141–55.

BIOGRAPHIES

RUI MAO received his B.S. (1997) and M.S. (2000) in computer science from the University of Science and Technology of China, and another M.S. (2006) in statistics and his Ph.D. (2007) in computer science from the University of Texas at Austin. After three years at Oracle USA Corporation, he joined Shenzhen University in 2010, where he is now an associate professor in the College of Computer Science and Software Engineering. His research interests include universal data management and analysis, and high-performance computing. He has about 50 publications, and his work on the pivot space model was awarded the SISAP 2010 Best Paper award.

HONGLONG XU received a B.S. degree in computer science from Shenzhen University in 2010, and is now a Ph.D. student in communication engineering at Shenzhen University.

WENBO WU received a B.S. degree in computer science from the University of Texas at Austin, and is now a Ph.D. student of statistics at the University of Georgia.

JIANQIANG LI received his B.S. degree in mechatronics from Beijing Institute of Technology, China, in 1996, and his M.S. and Ph.D. degrees in control science and engineering from Tsinghua University, Beijing, China, in 2001 and 2004, respectively. He worked as a researcher at the Digital Enterprise Research Institute, National University of Ireland, Galway, in 2004–2005. From 2005 to 2013, he worked at NEC Labs China as a researcher, and the Department of Computer Science, Stanford University, as a visiting scholar in 2009–2010. He joined Beijing University of Technology in 2013 as Beijing Distinguished Professor. His research interests are in Petri nets, enterprise information systems, business processes, data mining, information retrieval, semantic web, privacy protection, and big data. He has over 40 publications and 37 international patent applications.

YAN LI received his B.S. in electronic engineering and information science from the University of Science and Technology of China in 2001, and his Ph.D. from Paris 11 University in 2007. He joined Shenzhen University in 2008, and is now an associate professor in the College of Computer Science and Software Engineering. His research interests include low-noise signal processing and mixed signal IC design.

MINHUA LU received her B.S. in electronic engineering and information science from the University of Science and Technology of China in 2001, and her Ph.D. degree in biomedical engineering from the Hong Kong Polytechnic University, China, in 2007. She joined the Department of Biomedical Engineering, Shenzhen University in September 2007, and is now an associate professor and associate head of the department. Her main research interests include biomedical ultrasound imaging, medical instrumentation, tissue elasticity imaging, and image processing. She has published more than 60 technical papers, and holds 3 invention patents.

Bayesian Coalition Game for the Internet of Things: An Ambient Intelligence-Based Evaluation

Neeraj Kumar, Naveen Chilamkurti, and Subhas C. Misra

Neeraj Kumar is with Thapar University.

Naveen Chilamkurti is with LaTrobe University.

Subhas C. Misra is with the Indian Institute of Technology Kanpur.

ABSTRACT

With the evolution of the Internet and related technologies, there has been an evolution of new paradigm, which is the Internet of Things, IoT. In the IoT, a large number of objects/devices on the Internet are connected to one another for information sharing, irrespective of their locations. These devices may be interconnected with one another using various network protocols and standards to exchange information between them. The underlying network used for information exchange generally has built-in intelligence, which is called ambient intelligence, so that it can make adaptive decisions for information exchange between these objects in theh IoT. This article provides a performance evaluation of the Bayesian coalition game among these objects in the IoT environment by using the concepts of game theory and LA. In comparison to the existing solutions, LA are assumed to be the players in the game having *variable learning rates* in the coalition game. Most of the existing solutions have considered constant learning rates of the players in the game, which may lead to the possibility of local optima at some points. Each player decides its actions using competitive learning, having *variable learning rates*, based on the newly defined *utility function*, which leads to the achievment of a Nash equilibrium in the game quickly. Each player receives feedback from the environment corresponding to the actions taken in a unit interval of time. The performance of the proposed scheme was evaluated with respect to various performance evaluation metrics. The results obtained show that the proposed scheme is useful in the IoT environment.

INTRODUCTION

With the evolution of the Internet and related technologies, a large number of objects/devices on the Internet are connected to one another for mutual resource sharing and communication. According to a recent survey [1, 2], by 2020, the ratio of the number of devices connected to the Internet to the number of persons will be between 2–3. These devices communicate with one another using the existing state-of-the art protocols and technologies that provide communication between remote objects such as smart sensors and actuators, vehicles, PDAs, and smart machines [1]. Due to the ease of interaction between these devices using the underlying technologies and protocols, the Internet of Things (IoT) finds many applications in data sharing and knowledge transfer to provide facilities such as seamless connectivity and resource sharing to end users. With an aim to provide large numbers of services to the end users on the fly with low operational and maintenance cost, the IoT is expected to become next generation technology in which intelligent objects/devices can make decisions on their own and respond as per the directions from remote places other than the Internet [1].

The devices connected to the IoT may be of heterogeneous types with respect to their operational mode, services provided by them, and communication technologies. Thus, connecting all these devices and then accessing them is one of the biggest challenges in this environment. As these devices may be located at different places across the globe connected to the Internet, the rate of data transfer, security issues, and mode of transfer decide the overall performance of any solution proposed for such an environment. The devices connected to an IoT may use various forms of short- and long-range communication for data sharing among themselves. Short-range communication includes WiFi, RFID, and Bluetooth; long-range transmission includes WiMAX, Long Term Evolution (LTE), and cellular communication. Also, devices may also be able to communicate with or without centralized control, as they have built-in intelligence to make independent decisions on their own in this environment [2–4].

Although the IoT is envisioned to become the next generation technology for providing seamless connectivity to a large number of devices on the Internet, there are several challenges that need to be addressed before fully

0163-6804/15/$25.00 © 2015 IEEE

accessing this technology. Normally, the devices connected to the IoT have low battery and computation power, so how the operations are performed keeping the constraints of minimum battery and computation power is one of the challenges in the IoT environment. Also, the types of communication protocols and infrastructure support needed for providing communication between different devices is an important issue. Apart from these, as these devices generate large volumes of data, the method using which this data can be stored and retrieved from anywhere is another big issue that requires special attention.

Keeping focus on the above constraints and challenges in an IoT environment, in this article, we provide a scheme for data sharing between these devices using the concepts of the Bayesian coalition game (BCG) and *networks of learning automata* (LA). We have assumed that networks of LA are stationed at the devices, which form coalitions among themselves using conditional probability. The relationship among the devices is represented using an acyclic directed graph (ADG). LA are assumed to be the players in the game stationed at devices, which perform a finite number of actions such as data capturing and communication with other devices situated at either local or global sites. Corresponding to each action performed by the automaton, it may receive a reward or a penalty from the environment, and accordingly update its action probability vector. To make a balance between the moves of the players in the game, Nash equilibrium (NE) may occur, in which no player can take any undue advantage of the moves of the other players in the game. After a finite number of iterations, the game stabilizes toward an optimal solution.

The rest of the article is organized as follows. The next section illustrates the background and preliminaries on BCG and LA. The following section illustrates the proposed approach with the communication infrastructure required for communication between different objects in this environment. The fourth section provides an evolution of the proposed scheme with respect to various performance evaluation parameters. The final section provides the conclusion and future directions for this work.

Background and Preliminaries

Figure 1 shows how various objects/LA are deployed in the IoT environment and then interact with one another for data sharing. LA may be deployed at either the local site or global sites connected to each other using wired/wireless connections. LA take input parameters and act according to these parameters to produce an output after interaction with the environment, which in this case is the IoT, consisting of diverse types of objects such as vehicles, actuators, sensors, and smart devices. LA execute a learning algorithm to make a decision about which action should be taken to fulfill their goals. LA improve their solutions in each interaction after receiving feedback from the IoT environment. LA select an optimal action from a number of actions performed through a series of interactions with the environment [5–14]. The prime objective of each automaton is to find the optimal solution with minimized penalty from the environment [11]. Following are the components of LA and the coalition game.

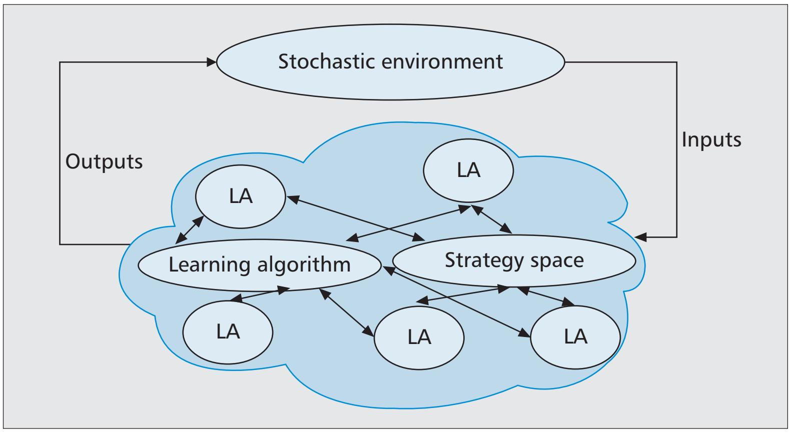


Figure 1. LA interaction with an environment.

Learning Automata as the Players

An automaton is a code executed to perform an action using some predefined rules and strategies. These predefined rules and regulations can be defined in an algorithmic form, which may be executed at a centralized or distributed place in an environment where that code resides. Hence, a central theme of the design of any solution using LA concepts is the placement of this code and its execution to get the results. This code may also be treated as an object that can perform multiple actions and collaborate with other objects to achieve its goals. As shown in Fig. 1, multiple automata (objects) collaborate with one another and interact with their environment to produce an output. These objects can share their data to execute various functions to accomplish their tasks.

Mathematically, LA can be defined as 5-tuple with attributes as $<Q_1, K_1, P_1, \mu, H>$, where $Q_1 = \{q_1, q_2, ..., q_n\}$ is a finite set of states of an automaton, $K_1 = \{k_1, k_2, ..., k_n\}$ is a finite set of actions performed by it, $P_1 = \{p_1, p_2, ..., p_n\}$ is a finite set of responses received from a given environment, $\mu : Q_1 \times P_1 \rightarrow Q_1$ maps the current state and input from the environment to the next state of the automaton, and H is a function that maps the current state and inputs from the environment to the state of the automaton [5–14]. The environment is the place where all these objects interact with each other to do their tasks. In this process, these automata also update the action probability vectors. To get the best results from any solution, there is a requirement of collaboration between different LA working in the IoT environment. Hence, LA in the proposed solution are assumed to be the players that collaborate with one another and interact with the IoT environment.

Learning Environment

As discussed above, LA operate in a place called the environment and execute the learning algorithm to complete their tasks. The environment in our current solution is the IoT where large

A bargaining game is formulated among the players of the game showing the moves with respect to the conditional probability among the different actions taken by the players. The conditional probability is defined by Bayes TTheorem.

numbers of objects are interconnected with each other. These objects (e.g., vehicles, sensors, actuators, and smart gadgets) vary in terms of their computing and processing power to create a heterogeneous environment.The LA continuously interact with the environment and perform their actions after taking the reinforcement signal from the environment. In the current solution, we have assumed that LA act as the players in the coalition game. The initial action taken by all the players are random as there is no feedback available from the environment at that time. But as the game progresses, each player carefully selects its moves after taking the cooperation with respect to the moves of the other players.

Mathematically, the environment can be defined as a triplet $<A, B, C>$, where $A = (A_1, A_2, \ldots, A_n)$ is a finite number of inputs, $B = (B_1, B_2, \ldots, B_n)$ is a set of values for the reinforcement signal, and $C = (C_1, C_2, \ldots, C_n)$ is a set of penalty probabilities associated with A_i, $1 \leq i \leq n$. After taking the initial actions randomly, each player interacts with the environment and receives its feedback in terms of a reward or a penalty from the environment. During this process, each player also updates its action probability vector, which it uses for all its future actions to be taken. The environment can give constant or variable feedback for each action taken by the players. In the former case, players learn from the environment with constant learning; in the latter, they learn with variable learning. In both cases, the action probability vector is updated with a learning parameter, which can be tuned during the experimental setup and influences the convergence of the designed solution in the IoT environment.

Action Probability Updates

According to the response received from the environment in terms of a reinforcement signal, which may be constant or variable, each automaton decides its next state to move on in a coalition game. Two sets are used for this purpose: one is used to keep track of all the strategies to be taken by the players in the game, and the other is to keep track of the various states in the coalition game. Corresponding to each input parameter and action, an automaton updates its action probability vector by using the learning algorithm. There are various schemes for updates of the action probability vector, but in the proposed solution, we have considered the Linear Reward-Inaction scheme (LR-I). In this scheme, if LA receive a reward from the environment, the action probability is updated according to the reinforcement signal received from the environment. Otherwise, the action probability vector remains the same for all the actions taken in the coalition game. The equations for the updates of the action probability vector are illustrated in [5–14].

The Proposed Approach

The proposed approach consists of coalition formation with merging and splitting, depending on the utility of the individual player in the game. A bargaining game is formulated among the players of the game showing the moves with respect to the conditional probability among the different actions taken by the players. The conditional probability is defined by Bayes' Theorem. Various actions of the proposed scheme are described as follows.

Coalition Formation

Before performing any operation in the environment where the players in the game are performing their individual actions, a coalition must be built. As the players in the game are devices connected to the Internet, these players can either transmit the packets to the next player or keep them until a suitable node is not found. In the latter case, these may act as relay nodes in the network. The state of each player in the game is either a transmitting mode or a relay mode. The state space consists of representation of transition probabilities from transmitting to relay or vice versa. Players take actions depending on the conditional probability of transitioning from one state to another in the coalition game. Let A_i be an action performed by the player, p_i^{trans} the probability of transmitting, and p_i^{relay} the probability of relaying the packets. Also, let $p(A_i)$ be the total probability of taking all the actions in the game. Then the conditional probability of performing an action by the players in the game can be computed by taking the ratio of p_i^{trans} to $p(A_i)$ subject to the successful execution of p_i^{relay}. Coalition among the players of the game is built based on the utility function UF_i computed by the total revenue generated to the total cost incurred in maintaining the resources in the coalition game.

Let the probability of transmission with respect to distance be $(p_i^{trans}(d_{ij}))$. Also, let the number of channels, n_{ch}, with respect to the maximum number of channels is denoted as $n^{max(ch)}$ and the maximum distance as d_{ij}^{max}. The distance is kept in computation of the utility function as it is used for measuring the probability of transmission from the source to the destination in a unit interval of time. For each transmission from the source to destination, revenue and cost are associated for successful delivery of packets. Revenue would be generated by bargaining the resources from the service provider, which in this case may be a cloud data repository from which various resources can be consumed by different devices in the IoT environment. Similar to the revenue, some cost is also associated for accessing the services from the service provider. The utility is constructed with respect to the overall revenue generation r and associated cost c. Utility is generated and maintained by each player for revenue generation and cost in the game. This utility generated is evaluated at regular intervals for taking adaptive actions by the players in the game. Let $S = (s_1, s_2, \ldots, s_n)$ be the strategy space from which a player kicks up one of the strategies for execution. The strategy chosen is dependent on the current state of the player in the game and the opponent position based on which each player selects its actions.

The coalition among the players of the game is built based on the conditional probability and evaluation of the utility function at regular inter-

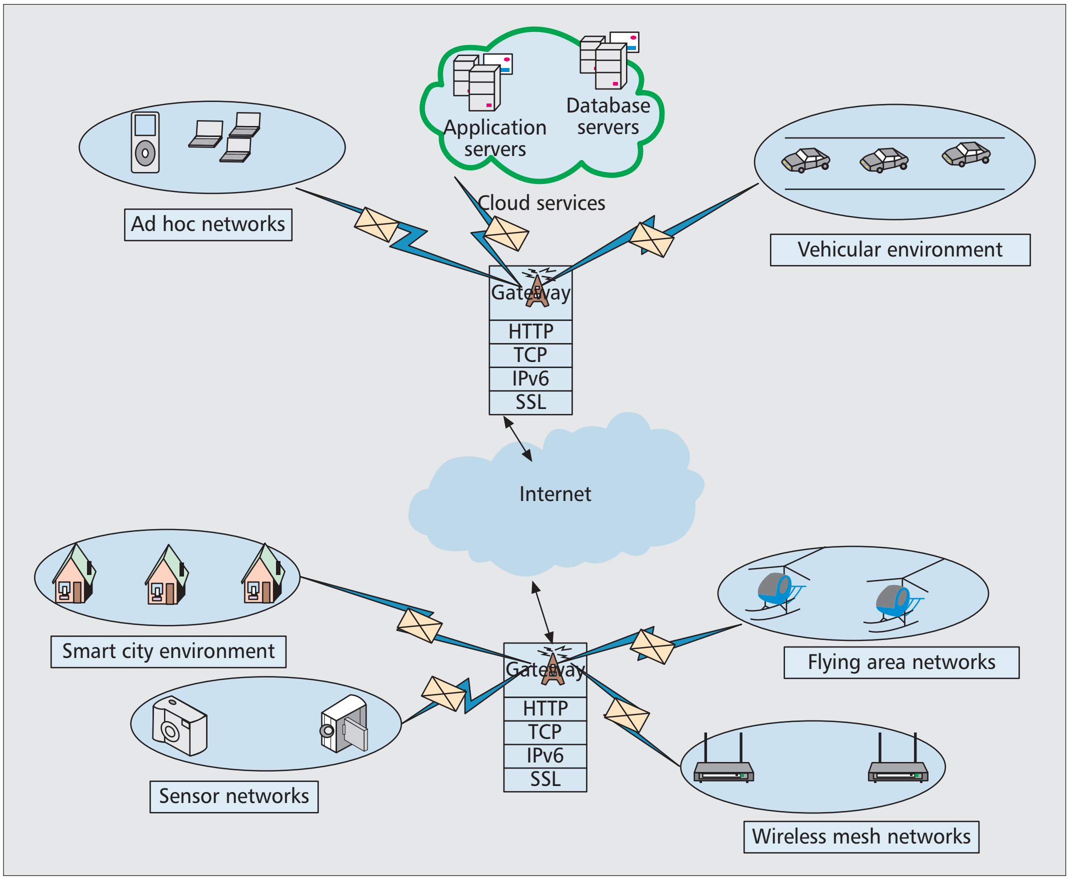


Figure 2. Generalized architecture in the IoT environment.

As the players in the game perform some actions, their utilities get changed due to the reward and penalties they get from the environment. Each player may stay in the current position by keeping the packets received from the previous nodes.

vals by the players. Initially, each edge of the ADG is associated with the attributes, including the identification of the player, *id*, strategy chosen, s_i, number of actions performed $n(A)$, number of rewards $n(r)$, number of penalties $n(p)$, and utility function id UF_i^{id} of the player. The initial utility of each player is computed in the beginning and updated as the players interact with the environment with constant rewards and penalties. All the strategies have equal probability of selection at the initial stage, but their probabilities change after interaction with the environment. As the players in the game perform some actions, their utilities get changed due to the reward and penalties they get from the environment. Each player may stay in the current position by keeping the packets received from the previous nodes. In this case, it acts as a relay node in the network.

Merging and Splitting of Coalitions with Variable Learning Rates

As discussed above, coalitions among players are formed based on the cost and revenue in the utility function with any object in the IoT environment. Each player has the intention to increase its own utility for maximum benefits, which is dependent on the number of awards and penalties it gets from various actions performed in a unit interval of time. Thus, initially, the probability of selection of all the actions is equal, but it may be increased or decreased based on the actions performed. The players in the game have the flexibility to move from one coalition to another depending on the current utility of the players. The strategy space of the players in the game may be variant or invariant with time. The strategy space remains the same irrespective of a player's move if it is time invariant. On the contrary, it is adaptive with respect to the moves of the players if it varies with time. A discrete time interval is considered for strategy adaptation with respect to the time in which each node computes its number of actions taken in comparison to the other nodes in the game. According to the moves taken by the other players in the game, each player selects one of the strategies mov^{forw}, mov^{back}, and *stationary*, where mov^{forw}, mov^{back}, and *stationary* are the strategies to move forward, move backward, or remain stationary in the current position.

Initially, all the coalitions are assumed to be of equal probability. Thus, players can join any coalition with random probability. However, as the number of moves are taken by the players in the game, their utilities may increase or decrease, and accordingly, the players decide the next action to be taken. As we have taken a variable learning rate in the proposed scheme for all the players in the game, depending on the utility of each player in the game, each player decides which action it should take by watching the moves of the other players in the game. Joining or leaving a coalition depends on the conditional probability with respect to the actions taken. The average packet delivery ratio with respect to strategies such as mov^{forw}, mov^{back}, and *station-*

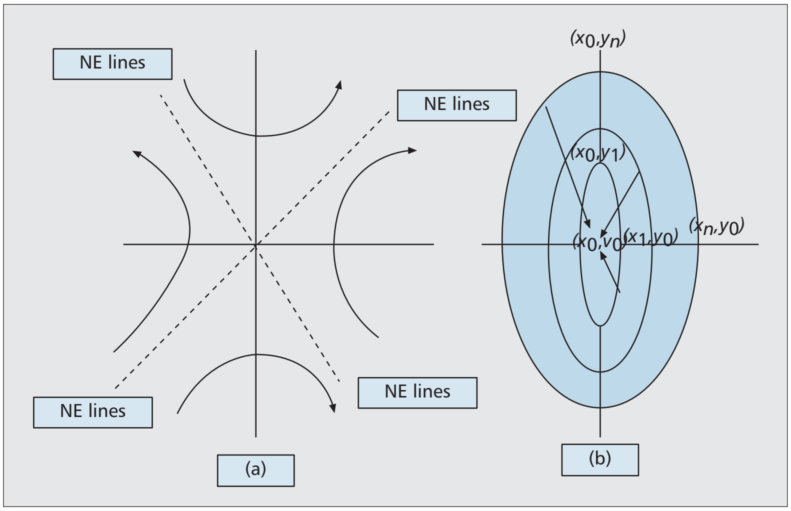


Figure 3. Achieving NE with variable learning rates.

ary, and the learning rate of the players is taken to set a threshold *thr* for making a decision by the players to stay in the current coalition or to move to the other. If utility function (UF) > *thr*, the players can join a new coalition, and the ADG representation of players and their actions will be split; otherwise, two or more coalitions are merged to form a single coalition.

Avoiding Local Minima with Variable Learning rates

The learning rate is an important parameter that decides the convergence of any solution in the coalition game. The learning rate may be constant or variable. In the case of a constant learning rate, all the players learn from the environment with a constant parameter, and in return, the environment also provides constant feedback in terms of a reward or a penalty to all the players with respect to the actions taken by them in the game. However, in the case of variable learning rates, the players learn with different learning rates, and the environment also provides different feedback to the players with respect to the actions taken by them. Hence, the variable learning rate may be considered a competitive learning process in which players have competition among themselves to learn more from the environment so that more rewards can be obtained from the environment, which results in faster convergence to an optimal solution. Most of the existing solutions in the literature, such as [2–8], have considered constant learning rates of the players in the game. At each iteration in the game, players can choose a pure or mixed strategy. A pure strategy is one in which the players select one of actions from the strategy space deterministically, while in the mixed strategy they may choose according to the probability distribution over the actions set. In both cases, each player has the intention of maximizing his/her UF function to gain maximum benefit in the game. Figure 3 shows how constant (a) and variable (b) learning rates result in faster convergence of an optimal solution to achieve NE in a multi-player game. Figure 3a shows how NE lines diverge from the point of convergence in the constant learning rates, while in Fig. 3b, with variable learning rates the point of convergence (x_0, y_0) is reached. There are various points represented by (x_i, y_i) in Fig. 3, which represent the moves of the players such that each player tends to move toward the convergent point (x_0, y_0).

Convergence/Divergence from the Optimal Value

As the players play the game, they may perform a number of actions that may lead to converging/diverting from an optimal solution. Let us consider a stochastic *n* players game, in which actions of the players are represented in an $n \times n$ matrix. Let α be the probability of taking the first action by the players row-wise in the matrix; then $1 - \alpha$ is the probability of taking the second action. Thus, α, $1 - \alpha_i$, and $1 \le i \le n$ are all pairs of possible actions taken by the players in the game row-wise in the $n \times n$ matrix. Similarly, we can define β, $1 - \beta_i$, and $1 \le i \le n$ as the actions to be taken column-wise. Hence, we can represent the actions of the players in an $n \times n$ matrix as $(\alpha_{ij}, \beta_{ij})$, $1 \le i \le n$, $1 \le j \le n$. Hence, for any pair of strategies from this matrix, we can compute how much is the UF with respect to the total cost and revenue generated in the game. We can define a combined strategy in an *n*-dimensional Euclidean space from the above defined matrix in which all the actions taken by the players are represented.

For any number of strategies from the strategy space, the players can get maximum or minimum UF value according to their actions performed in the game. The players can adjust their strategies at each iteration, which results in an increase or decrease of their UF value during different iterations in the game. The change in the UF value can be formulated by taking the partial derivative with respect to strategies selected from the strategy space. By equating the partial derivative to zero, we can get the number of states from the strategies space, and then again taking the differentiation and finding the points where the second derivative is positive, negative, or zero. If it is negative, we have a point of maxima, that is, the expected UF of the players will be maximum in that state. If it is positive, the expected UF of the players will be minimum. Otherwise, the expected UF will be constant. Hence, based on this criteria, each player decides whether it can move forward or backward, or stay in the current position in the game. At each iteration, each player compares its expected UF with the current UF, and if the difference between both the values is beyond a predefined threshold, say δ, the players will learn more by watching the moves of the other players so that NE can be achieved quickly in the game. In this case, players can increase their step size of learning for faster convergence. On the other hand, if players know that they are in a better in comparison to their opponents and have higher chances of getting maximum UF, the step size of the learning rate can be constant so that NE can be achieved. Thus, if all the players in the game follow this strategy of increasing or keeping their learning rates constant according to the situa-

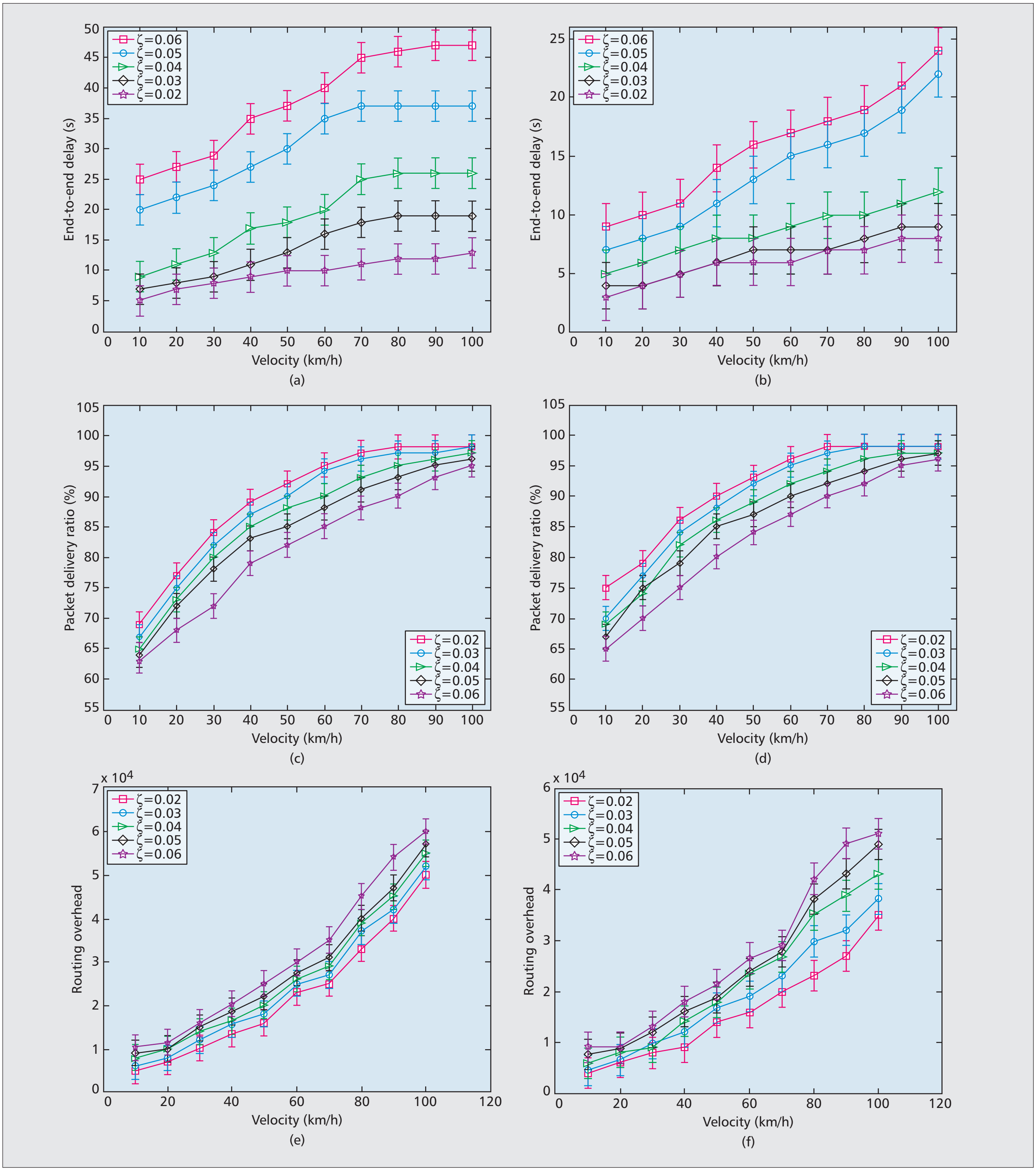


Figure 4. a) E2ED with beacon interval of 1 s; b) E2ED with beacon interval of 3 s; c) PDR with beacon interval of 1 s; d) PDR with beacon interval of 3 s; e) routing overhead with beacon interval of 1 s; f) routing overhead with beacon interval of 3 s.

tion, the game will be a balanced game, and the chances of getting the NE are greater.

As shown in Fig. 3, the constant learning rates of the players may lead to divert them from the optimal solution, which may result in a lesser probability of achieving NE points (Fig. 3a). But as the learning rates are varied according to situation and moves of the other players in the game, faster convergence toward an optimal solution can be achieved, thereby leading to NE in less time (Fig. 3b). As players can see the current and relative positions of the other players, it leads to a balance in the game from players' strategy selection perspectives, which ultimately controls the NE in the game, which is one of the primitives to analyze an n-player game in any environment.

Performance Evaluation

The proposed scheme was evaluated on the ns-2 platform with simulation for urban mobility (SUMO) [15] using the random waypoint mobility model. For demonstration of the proposed scheme in an IoT environment, we have specifically considered a vehicular network in which LA are assumed to be stationed on the vehicles with coalition formation among them using Bayes' Theorem. In this waypoint model, a total of 500 vehicles have been considered with variable speed between 10–100 km/h. Vehicles form clusters among themselves. The communication range threshold for the vehicles was selected as 200 m. The road is assumed to be divided into different clusters, with each cluster containing 100–400 vehicles. The following parameters are selected for evaluation of the proposed scheme:

> *The performance of the proposed scheme was evaluated with respect to various performance evaluation metrics based upon which it can be concluded that the proposed scheme can be useful for information exchange between different devices connected to the Internet.*

- *End-to-end delay (E2ED)*: It is defined as the total delay incurred for information transfer from source to destination.
- *Packet delivery ratio (PDR)*: It is defined as the ratio of successful transfer of packets to the total number of packets transferred.
- *Routing overhead*: It is defined as the additional packets transferred in addition to the normal routing packets.

The proposed scheme is simulated on 95 percent confidence interval with 20 simulation runs by taking the average values on these simulation runs. The results obtained are described in detail as follows.

Figures 4a and 4b show the variation of E2ED delay with beacon interval of 1 s and 3 s. We have varied the learning rates (ξ) of the players in the game from 0.02 to 0.06 in all the experiments conducted. As shown in the figure, with a variation in the velocity of the vehicles in the game, there is a variation in E2ED. As observed from the results, with an increase in the velocity and learning rates, the E2ED decreases. This is due to the fact that with the increase in learning rate, more data is available to LA to perform various operations; and during this process, LA select the best moves in the game using the UF; that is, the number of rewards is higher than the number of penalties in the game, which results in a decrease in the E2ED in the scheme. The increase in the beacon interval from 1 to 3 s results in an increase in E2ED, which can be observed in Fig. 4b.

Figures 4c and 4d show the variation in the PDR with variation in the velocity of the vehicles and learning rates from 0.02 to 0.06. It has been observed from the results obtained that with an increase in the learning rates, even at high velocity of the vehicles, PDR increases. This is mainly due to the reason that with an increase in the learning rate, the players in the game have higher probability of taking the correct moves, as they have large strategy space to select the best strategy. This results an overall increase in the PDR of the network, as observed in Figs. 4c and 4d. With the high velocity of the vehicles, there is a high chance of topological changes, but the players in the game select their destination by watching the moves of the other players. This results in a decrease in the penalty these players get from the environment and an increase in the UF. Hence, there is an increase in PDR with an increase in the learning rates, even with an increase in the velocity of the vehicles.

Figures 4e and 4f show the variation of routing overhead with variation in ξ from 0.02 to 0.06 and velocity of the vehicles. The routing overhead increased with an increase in the velocity of the vehicles along with the learning rate of the players in the game. This is expected, as the process of learning is slow, and good results are expected at slow learning rates. Moreover, due to the dynamic nature of the nodes in the network, quick decisions are expected. As the proposed scheme is based on the learning of the players to make adaptive decisions, so with an increase in beacon interval from 1 s to 3 s, it takes a longer time to make a decision about the final destination selection from the source/intermediate nodes. Hence, with an increase in velocity and learning rates, the routing overhead is increased, as observed in Fig. 4.

Conclusion

With the exponential growth in the usage of Internet-enabled devices, information exchange with respect to constraints such as longer E2ED, lower PDR between these devices has become a major challenge in recent times. As these devices may operate in a heterogeneous environment with varying functionalities of these devices, designing a protocol/standard for communication among these is an important issue. In this article, we have provided a performance evaluation of ambient intelligence scheme for information exchange between these devices/objects connected to the Internet using the concepts of BCG and LA. LA are assumed to be the players in the game, stationed at these intelligent devices and forming coalitions among themselves using conditional probability. A new utility function is designed for interaction within these intelligent devices. The value of the utility function increases or decreases based on feedback received from the environment. Varying the learning rates of the players in the game was selected for faster convergence of the solution. The performance of the proposed scheme was evaluated with respect to various performance evaluation metrics based on which it can be concluded that the proposed scheme can be useful for information exchange between different devices connected to the Internet.

In the future, we plan to optimize RFID tag deployment for collection of data from various objects in an IoT environment. The secure collection and distribution of RFID tags would also be analyzed.

References

[1] A. Zanella *et al.*, "Internet of things for Smart Cities," *IEEE IoT J.*, 2014.

[2] O. Bello and S. Zedally, "Intelligent Device-to-Device Communication in the Internet of Things," *IEEE Sys. J.*, 2014.

[3] N. Kumar and J. H. Lee, "Peer-to-Peer Cooperative Caching for Data Dissemination in Urban Vehicular Communications," *IEEE Sys. J.*, vol. 8, no. 4, 2014, pp. 1136–44.

[4] A. Dua, N. Kumar, and S. Bawa, "A Systematic Review of Routing Protocols for Vehicular Ad Hoc Networks," *Vehic. Commun.*, vol. 1, no.1, 2014, pp. 33–52.

[5] A. V. Vasilakos and G. I. Papadimitriou, "A New Approach to the Design of Reinforcement Schemes for Learning Automata: Stochastic Estimator Learning Algorithm," *Neurocomputing*, vol. 7, no. 3, 1995, pp. 275–97.

[6] J. A. Torkestani, and M. Meybodi, "Mobility-Based Multicast Routing Algorithm for Wireless Mobile Ad-Hoc Networks: A Learning Automata Approach," *Computer Commun.*, vol. 33, no. 6, 2010, pp. 721–35.

[7] A. V. Vasilakos and G. . Papadimitriou, "A New Approach to the Design of Reinforcement Schemes for Learning Automata: Stochastic Estimator Learning Algorithm," *Neurocomputing*, vol. 7, no. 3, 1995, pp. 275–97.

[8] S. Misra, V. Krishna, and V. Saritha, "LACAV: An Energy Efficient Channel Assignment Mechanism for Vehicular Ad Hoc Networks," *J. Supercomputing*, vol. 62, no. 3, 2012, pp. 1241–62.

[9] N. Kumar and J. Kim, "ELACCA: Efficient Learning Automata Based Cell Clustering Algorithm for Wireless Sensor Networks," *Wireless Pers. Commun.*, vol. 73, no. 4, 2013, pp.1495–1512.

[10] N. Kumar, N. Chilamkurti, and J. P. C. Rodrigues, "Learning Automata-Based Opportunistic Data Aggregation and Forwarding Scheme for Alert Generation in Vehicular Ad Hoc Networks," *Computer Commun.*, vol. 59, no. 1, 2014, pp. 22–32.

[11] N. Kumar and N. Chilamkurti, "Collaborative Trust Aware Intelligent Intrusion Detection in VANETs," *Computers and Electrical Engineering*, vol. 40, no. 6, 2014, pp. 1981–96.

[12] P. V. Krishna *et al.*, "Virtual Backoff Algorithm: An Enhancement to 802.11 Medium-Access Control to Improve the Performance of Wireless Networks," *IEEE Trans. Vehic. Tech.*, vol. 59, no. 3, 2010, pp. 1068–75.

[13] S. Misra *et al.*, "Random Early Detection for Congestion Avoidance in Wired Networks: A Decentralized Pursuit Learning Automata Like Solution," *IEEE Trans. Sys. Man Cybern*. vol. 40, no. 1, 2010, pp. 66–76.

[14] A. F. Atlasis, N. H. Loukas, A. V. Vasilakos, "The Use of Learning Algorithms in ATM Networks Call Admission Control Problem: A Methodology," *Computer Networks*, vol. 34, no. 3, 2000, pp. 341–53.

[15] sumo-sim.org/userdoc/Tutorials.

Biographies

Neeraj Kumar is working as an associate professor in the Department of Computer Science and Engineering, Thapar University, Patiala. He has received his M.Tech. from Kurukshetra University, Haryana, followed by his Ph.D. from SMVD University, Katra, in computer science and engineering. He was a postdoctoral research fellow at Coventry University, United Kingdom. He has more than 100 research papers in leading journals and conferences of repute. His research is supported by UGC and TCS.

Naveen Chilamkurti is currently acting head of the Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria, Australia. He is also the inaugural Editor-in-Chief of the *International Journal of Wireless Networks and Broadband Technologies* launched in July 2011. He is currently serving as a Technical Editor for *IEEE Wireless Communications* and an Associate Technical Editor for *IEEE Communications Magazine*. He has published about 165 journal and conference papers. His current research areas include intelligent transport systems, wireless multimedia, wireless sensor networks, and so on. He is also an Associate Editor for Wiley IJCS and SCN.

Subhas Misra is working as a professor in the Department of Industrial and Management Engineering at the Institute of the Indian Institute of Technology (IIT) Kanpur. He received his Ph.D. from Carleton University and P.D.F. from Harvard University. He has experience in both academia and industry. Before joining IIT Kanpur, he worked for several years in different companies, including Indian Telephone Industries Ltd (India) and Nortel Networks (Canada). He has 80 peer reviewed papers in reviewed journals from leading publishers. He has 18 prestigious awards for his academic recognitions, including a Research Excellence Award (United States), an Achievement Award (United States), Best/Outstanding Paper Award (United States), NSERC Fellowship (Canada), OGS (Canada), and Young Scientist Award. He is on the Editorial Boards of 12 prestigious international journals.

LDPA: A Local Data Processing Architecture in Ambient Assisted Living Communications

Kun Wang, Yun Shao, Lei Shu, Guangjie Han, and Chunsheng Zhu

Kun Wang and Yun Shao are with Nanjing University of Posts and Telecommunications.

Lei Shu is with Guangdong University of Petrochemical Technology.

Guangjie Han is with Hohai University.

Chunsheng Zhu is with the University of British Columbia.

ABSTRACT

In ambient assisted living, one of the most concerning problems is how the living status of the elderly can be accurately judged through the data collected by ambient sensors. To solve this problem, environmental influences should be sufficiently considered. Some researchers have proposed collecting data comprehensively by distributed environmental sensors, but how the massive collected data were to be analyzed and transmitted was ignored. In this article, we propose a local data processing architecture on a local server to analyze collected data. In this three-layer architecture, the latest received data is stored in a data gathering layer. Afterward, a data filtering layer checks the efficiency of data. Also, this layer classifies the received data into static data reflecting sensors' status and a real-time data stream reflecting quality of life. For static data, they are directly stored in a database, and the real-time data stream is divided into different levels. Based on these levels, a data analyzing layer reorganizes data into a neighborhood structure, which will be called RDAA. A risk factor is returned by RDAA and only abnormal data will be sent to a health care provider when its risk factor is larger than a given threshold. LDAP disperses the stress of remote centralized processing and data storage, which decreases the workload of the remote health care provider. Meanwhile, it also reduces network load and improves processing speed.

INTRODUCTION

Ambient assisted living (AAL) enhances the independent living ability of the elderly through various intelligent services, reducing the need for direct caregiving indoors or outdoors. For example, living status and rest regularity can be reported to a remote health care provider (RHCP) through several sensors in a house. Recently, this kind of intelligent services has obtained further application with the development of sensing and processing techniques. Examples of applications provided by AAL include sending alerts to elderly people to take medicine, monitoring living status to avoid dangerous situations, and detecting and analyzing physiological features to ensure health and safety. AAL technologies can also be applied to real-time communications that keep the elderly in touch with their families.

From the technical aspect, even though the main technologies of AAL are mature, several problems are exposed for further applications. As comprehensive observation is needed for the elderly, relevant tasks are always completed by distributed sensors [1], as shown in Fig. 1. Also, these distributed sensors need to work for a long time. Under these circumstances, however, the workload of the RHCP and network load will increase dramatically if all sensors transmit the collected data to the RHCP for centralized processing.

Specifically, due to the influence of different environmental factors, a single sensor can hardly meet the overall needs. For example, the elderly may carry brain sensors reflecting their brain activities, heart rate sensors protecting them from arrhythmia, and human body temperature sensors. Also, accurate judgment of living and physiological status can be disturbed by weather, temperature, transportation, state of motion, and exercise, so we need to deploy different sensors into the environment. These sensors generate massive private data, which can only be accessed by personal doctors or an RHCP remotely. If these massive data converge to an RHCP for centralized processing, the pressure on the RHCP is increased.

To this end, we mainly focus on rapid data processing and abnormal status detection to reduce the workload of RHCPs in distributed sensors of AAL. The contributions of our article are summarized as follows:

- In our local data processing architecture (LDPA), we propose a new concept of risk function, which provides the idea of quantifying the result of data analysis in AAL. Also, a risk threshold is introduced for comparison of quantified output.
- In LDPA, a new classifier with a learning mechanism is designed to simplify input. Specifically, this filtering mechanism is to store static data into a database and obtain efficient real-time input of risk function.

0163-6804/15/$25.00 © 2015 IEEE

- In LDPA, we propose a new reduced variable neighborhood search (RVNS)-based data analysis algorithm (RDAA) by applying RVNS to the data analysis of distributed sensors in AAL. Generally, RVNS is used in a combinational optimization problem with a fixed terminal condition [2]. However, we design a new terminal condition in RDAA based on risk function.

The rest of the article is organized as follows. In the following section, we present the related work of data processing in AAL. After that, we introduce RVNS algorithm. The three-layer architecture of LDPA is described in detail in the section on gathering, filtering, and analyzing data. Performance evaluation of the proposed RDAA is demonstrated after that. The final section concludes the article and gives our future research consideration.

Related Work

In AAL, it is necessary to collect data reflecting the surrounding environment of the elderly to obtain accurate and credible results. In terms of environmental data collection and accurate data analysis, related research can be summarized into two categories.

Reducing the Influence of Environmental Factors through Accurate Analysis

The authors of [3] proposed a system with self-configuration according to the surrounding environment. This system can improve the accuracy of data collected by sensors deployed at important positions indoors and GPS equipment outdoors. Through comprehensive records, more evidence can be collected for accurate analysis. The authors of [4] proposed an idea of sensing data being sent to different RHCPs for analysis based on their different experience and understanding. More detailed discussion can be launched when different exports are enrolled, providing more accurate analysis. The authors of [5] ensured the accuracy of collected data through the verification of a multi-layer model. Although these research papers take the accuracy of analysis into consideration, they ignore massive data transmission and the workload of the RHCP.

Obtaining Credible Data by Environmental Cognition

The authors of [1] took environmental factors into consideration. Through recognizing these factors, more accurate and comprehensive data can be collected. The authors of [6] proposed a system to detect abnormal behavior by recognizing the surrounding environment. This system adopts a real-time checkout mechanism, which is different from regular abnormal behavior detection. The authors of [7] proposed a distributed sensor system in the environment to provide overall detection. In this system, sensors deployed in the environment collect data to comprehensively reflect the living status of the elderly. However, these research papers did not propose any efficient analysis algorithm for data processing.

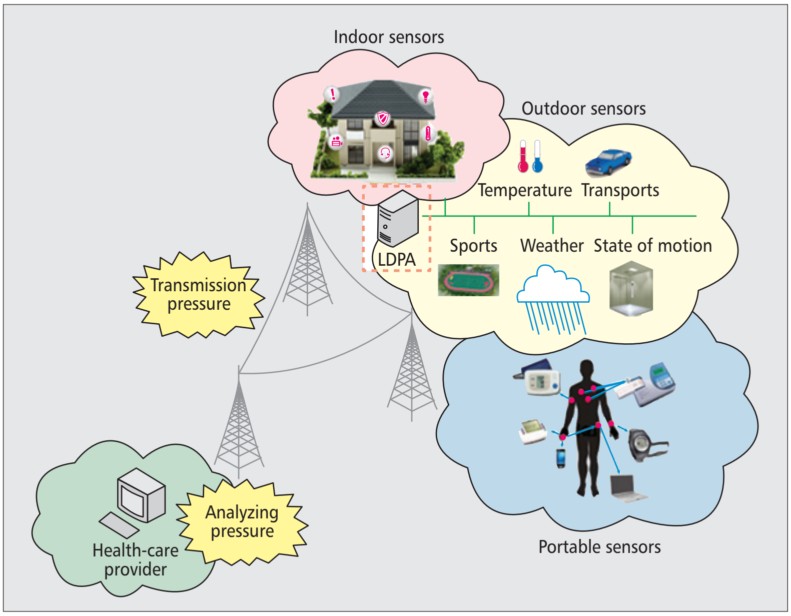


Figure 1. Problem statement of distributed sensors in AAL.

In terms of data processing architecture in AAL, the authors of [8] proposed a data collecting system based on wireless sensor networks and a mobile platform. In this system, portable sensors carried by the elderly can send physiological data to a processor that analyzes these data through a classifier. However, this system does not consider the influences of environmental factors. For example, judgment may be wrong if the elderly are exercising. The authors of [9] proposed a comprehensive monitoring architecture based on a network. It sends collected environmental data to the RHCP. This architecture concerns the deficiency of hospital monitoring when comprehensive data can be hard to achieve. However, massive data is sent to the RHCP for analysis, increasing the workload. The authors of [10] tried to build a smart house through the synchronization of sensors and behavior cognition. However, there are many unsolved problems in the fields of data transmission and analysis in AAL for several factors to be considered [11]. In addition, most data processing algorithms in AAL are based on conventional first-in first-out (FIFO), where data wait in a FIFO queue [12]. When processor speed can meet the needs of data, algorithm efficiency can still be ensured. However, when data increases dramatically, FIFO will no longer satisfy the need.

According to the description above, we propose a LDPA, mainly concerning data gathering, filtering and analyzing. This architecture shares the processing pressure of RHCP and reduces network load.

An Overview of the RVNS Algorithm

In this section, we introduce RVNS and its original algorithm, variable neighborhood search (VNS). VNS is a meta-heuristic algorithm for solving combinational optimization problems [2]. The whole execution process of VNS mainly includes two parts: constructing changeable neighborhoods systematically and searching for a locally optimal solution. The merits of VNS are reducing the calculation complexity through local search and avoiding local optima through changing neighborhood structure systematically.

In VNS, local search calls a subroutine to gradually optimize a local optimal solution in a neighborhood, and then reconstructs or switches neighborhoods according to this local optimal solution. The increasing calculation complexity of this subroutine call is time-consuming. Considering the necessity of rapid reaction toward abnormal situations in AAL, we propose a data analyzing algorithm (i.e., RDAA) based on RVNS in this article. As a simplified version of VNS, RVNS removes the local search, and chooses a random point, an element in the current neighborhood, as a local optimal solution. Then RVNS reconstructs or switches neighborhoods in reference to this random point, since RVNS simplifies calculation complexity, which ensures the rapid reaction of RHCP.

In RVNS, there is a solution space $S = \{x_1, x_2, x_3, \dots, x_n\}$, where initial solution x is obtained. For optimization problems, if $\forall x^* \in S$ satisfies $f(x^*) \geqslant f(x)$, solution x^* is regarded as a new feasible global optimal solution and replaces the old global optimal solution x, where $f(x)$ denotes a utility function. Correspondingly, for a global optimal solution x, there is a neighborhood structure $N(x)$ based on x. Just as $x \in S$, $N(x)$ is a subset of S ($N(x) \in S$). In a neighborhood structure with k partitioned neighborhoods, these specific neighborhoods are denoted as N_k, where $k \in [1, k_{max}]$. Particularly, when x is a feasible global optimal solution, these k specific neighborhoods are denoted as $N_k(x)$. N_k generates different neighborhood structures according to x through a series of matrix transformations.

Different from global optimal solution x^* (global maximum reached by optimization), a local optimal solution is denoted as x' ($x' \in S$, a local maximum reached by optimization). That is, we cannot find another x that satisfies $f(x') < f(x)$. In RVNS, x' is generated randomly in $N_k(x)$. When x' satisfying $f(x') > f(x)$, the value of k is returned, and the algorithm reconstructs the neighborhood structure according to x'. Otherwise, it switches to next neighborhood.

The description above presents an idea of partitioning in RVNS. The authors of [13] proposed an ACQE algorithm to optimize power allocation in smart grid, in which the solution space is divided into two sub-groups. However, the difference between algorithms like ACQE and RVNS is that the former generates different optimal solutions according to various combinations of sub-group divisions, while the global optimal solution RVNS finds solutions that remain unchanged during the reconstruction of groups' organizations (i.e., neighborhood structures). In addition, RVNS allows more flexible methods to divide the whole solution space into several neighborhoods.

Currently, RVNS-based meta-heuristic algorithms are always adopted to the location routing (LRP) [2]. Therefore, this is a new attempt to introduce this algorithm to data processing and analysis in AAL. The proposed RVNS-based analysis algorithm can quickly recognize information of abnormal status among massive data.

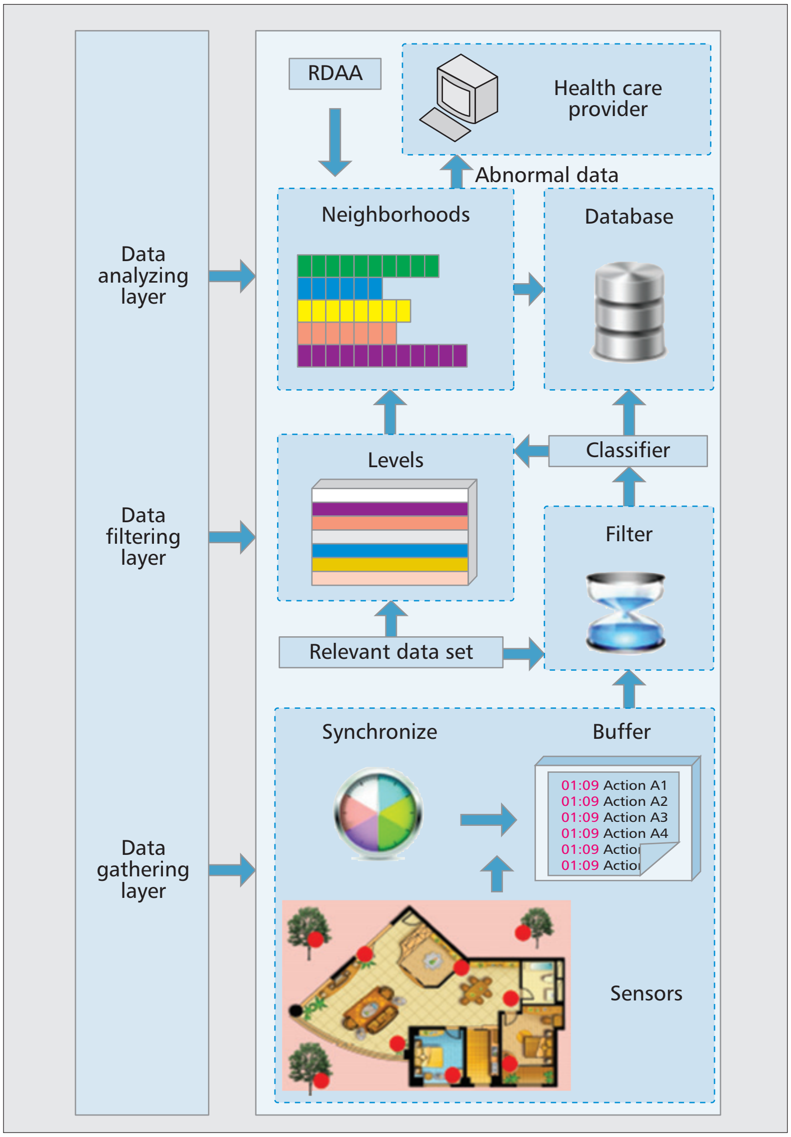


Figure 2. Structure of LDPA.

Local Data Processing Architecture

In this section, LDPA is described in detail. As shown in Fig. 3, LDPA can be divided into three layers: the data gathering layer (DGL), data filtering layer (DFL), and data analyzing layer (DAL). In DGL, a local server receives and stores relevant sensing data. Afterward, directly called by RDAA, these stored data are processed to meet the needs of RDAA in DFL. DAL will adopt RDAA to analyze these processed data through neighborhood reconstruction. RDAA calculates risk factor according to a complex risk function, which serves as a utility function. If the risk factor is higher than a given risk threshold, RHCP will receive an alarm signal with relevant data. Considering the characteristics by which RVNS can select a random point during local search, RDAA can improve analyzing efficiency with massive data and a complex risk function.

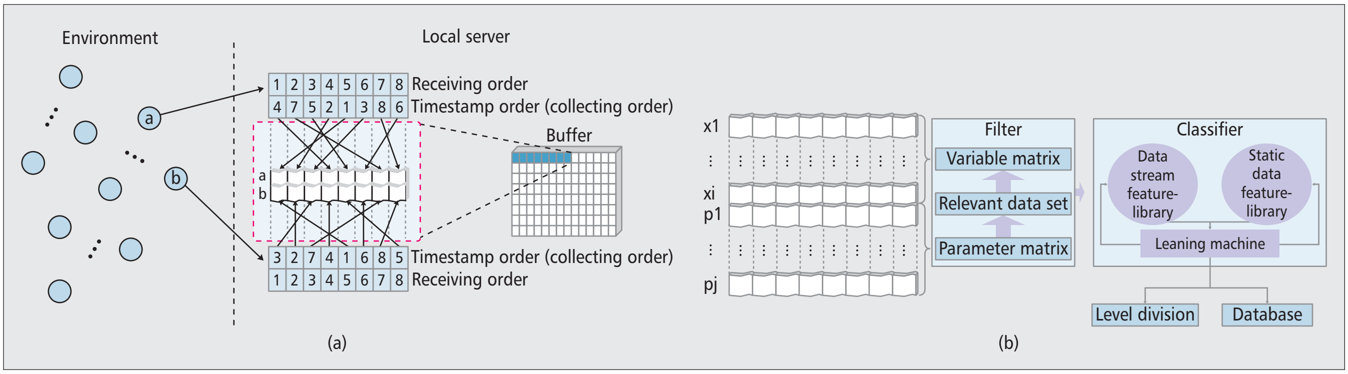

Environment
Local server
a
b
1 2 3 4 5 6 7 8 Receiving order
4 7 5 2 1 3 8 6 Timestamp order (collecting order)
Buffer
a
b
3 2 7 4 1 6 8 5 Timestamp order (collecting order)
1 2 3 4 5 6 7 8 Receiving order
(a)
x1
xi
p1
pj
Filter
Variable matrix
Relevant data set
Parameter matrix
Classifier
Data stream feature-library
Static data feature-library
Leaning machine
Level division
Database
(b)

Figure 3. a) Data gathering layer; b) data filtering layer.

Figure 2 presents the three-layer structure of LDPA. Sensing data are collected and stored in DGL, where red points denote the sensors deployed in the house and outside to collect relevant data to the local server. Also, some portable sensors carried by the elderly also send data (e.g., location, velocity, and psychological status). These data are stored in a buffer after synchronization in chronological order. Then data go through DFL. In this layer, efficient input of risk function is selected through a filter and classifier, by which part of the data (i.e., static data) is stored in the database, and the other part (i.e., data stream) is divided into different levels according to a relevant dataset. In DAL, these data are set in different neighborhoods according to their levels, and are called by RDAA to discover abnormal data. Finally, these abnormal data are sent to RHCP. When the buffer is filled up, a buffer cleaning mechanism is called to transfer the oldest data to the database.

Data Gathering Layer

In DGL, a distributed sensor architecture is deployed in the living environment of an elderly person. Data collected by these sensors are converged locally for processing. Different from RHCP, this local server is a computing and storage system that can communicate with RHCP. RHCP can control and visit a local server when necessary.

The deployed sensors can be divided into two categories. One is primary sensors, mainly consisting of portable sensors. These sensors are carried by an elderly person during her daily activities to monitor the psychological status and moving status, heart rate, velocity, height, and so on. However, these data cannot reflect risk accurately due to some environmental factors. For example, transportation can return abnormal velocity, and exercise can cause abnormal heart rate. Accordingly, several auxiliary sensors are needed in the living environment.

Initialization of these sensors is different according to environmental conditions. They cannot monitor the elderly all the time, but record meaningful data when the elderly are in relevant environments. Some important messages can be exchanged between primary and auxiliary sensors if necessary. All sensors set up communication with the local server to be used for data transmission and synchronization.

The receiving and storing processes are presented in Fig. 3a. Primary sensors and auxiliary sensors send the latest data to the local server periodically. Since the receiving order may vary greatly from the collecting order in reality, a timestamp is given, serving as the foundation of data realignment. The order of timestamping should be exactly the same as the collecting order. DGL will realign these data according to this timestamp, and put them into the relevant slot. In every slot, there is space for each sensor so that data from sensor a and b can be stored in the corresponding space, as shown in the red box (i.e., a magnified logical structure view of the buffer) in Fig. 3a. Specifically, recall the brain sensors mentioned in the introduction as an example. Due to the large amount of data needed in detecting brain activities, the task is always completed by a group of sensors. These sensors transmit the data to the server, where data is realigned according to timestamps, which synchronizes the data from the same sampling time.

In a buffer, many similar slots constitute a section, and several sections form the whole buffer. The buffer in a local server can be regarded as a twice-division structure logically, which is similar to a 2D array. In the first division, the whole buffer is divided into several sections. For example, if a buffer can totally store the data within one day, every section may store the data within four hours. Different slots are divided during the second division in each section. As for this example, there are 240 slots in a section to store the data received each minute.

In a local server, the receiving order of data is not always the collecting order as network condition varies from time to time. Accordingly, once data have been received, they are realigned according to timestamp. Only data with the same timestamp can be stored in a slot. Data with earlier timestamps are stored into prior slots. In every slot, space has been allocated for each sensor. When data from all sensors at a specific time have been stored, they enter DFL.

In addition, if all buffer spaces have been filled up, some data have to be dropped. In this article, a buffer cleaning mechanism is adopted to transfer the earliest data into the database.

Data Filtering Layer

Filter — Through DGL, relevant data have been stored in the buffer. Among these data, the sensing data collected by primary sensors form a VariableMatrix of risk function, denoted as S.

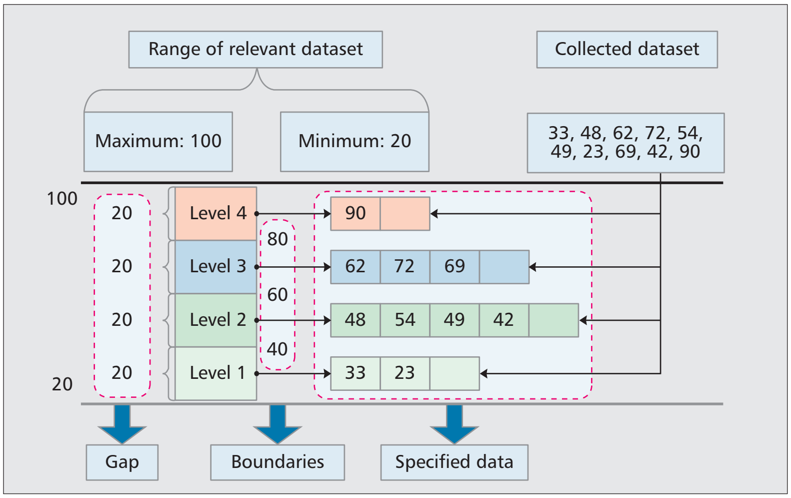


Figure 4. An example of level division.

Correspondingly, data from auxiliary sensors consist of a ParameterMatrix, denoted as *P*. The rows of *S* and *P* represent the data from the same primary sensor and auxiliary sensor at different sampling times, while the columns of *S* and *P* represent data from different primary sensors and auxiliary sensors at the same sampling time, respectively.

The inputs of risk function (described in DAL) can be achieved by combining VariableMatrix *S* with ParameterMatrix *P*, which consists of sensing data from different sensors collected at the same time.

It is noted that absolutely accurate data only appears in an ideal situation. In reality, input may mix up some inaccurate data due to network conditions. As for these data, we add a filter here to judge data accuracy. Meanwhile, a relevant dataset is introduced as a criterion. This relevant dataset stores a large amount of historical data collected by primary and auxiliary sensors, and RHCP updates this dataset periodically. As shown in Fig. 3b, the filter selects representative data in *P* at the same sampling time, and efficient ranges can be found in a relevant dataset. According to efficient ranges, some invalid data in *S* can be ruled out. To some extent, this filter can alleviate the interference caused by invalid data.

Classifier — In addition, sensors always return some data reflecting their work conditions. In this case, data collected in DGL can be divided into static data and data stream monitoring the living status of the elderly. If both kinds of data are served as input on risk function, risk function complexity increases and buffer space is taken up. To solve this problem, we design a classifier in DFL. It includes a data stream feature library, a static data feature library, and a learning machine. In comparison with a feature library, data stream and static data can be separated. In terms of static data, a classifier transfers them from buffer to database directly, serving as the data source of maintenance. Also, their buffer spaces are released. On the other hand, the data stream serves as the input of risk function. Specifically, *S* is a solution space of RDAA (described in DAL), while *P* provides relevant parameters to the solution space. Besides, feature libraries can be updated periodically to enhance analysis accuracy continuously due to the existence of the learning machine.

Level Division — Afterward, taking a relevant dataset as a criterion, data in the solution space are divided into several levels according to a group of data from a specific primary sensor. Since all data in the solution space are seen as the input of risk function, the algorithm accuracy of RDAA is not varying if different groups of data are used. Level division can be completed according to the maximum and minimum data groups in the relevant dataset. First, taking the number of levels needed, the gap of each level can be achieved by (maximum-minimum)/number of levels. Then the boundary of each level can be built based on the gap value and the minimum data group in the relevant dataset. Lastly, each of the collected data in the group can be specified into a corresponding level according to these boundaries. Figure 4 shows an example of level division. Boundaries of levels are set up according to the relevant dataset, and collected data are specified into different levels based on boundaries.

As a result, for the data stored in every slot, they are divided into a corresponding level, serving as the foundation of neighborhood construction.

Data Analyzing Layer

In this section, we demonstrate how data are analyzed by RDAA. First, we introduce risk function. Afterward, we discuss the details of the data analyzing process of RDAA. Simulation results are presented by comparison with FIFO.

Risk Function — Risk function is a complex function reflecting dangerous levels of behavior or psychological status in a specific environment, denoted as R. In risk function, variables are data from primary sensors, and parameters are data from auxiliary sensors.

Risk function is based on clinical data and the influence of different elements on the environment. It tries to evaluate the influence of the environment on the elderly through environmental parameters collected by auxiliary sensors, and then judge the data gathered by primary sensors. For example, heart rate sensors report the heart rate of an elderly person in an AAL system to judge arrhythmia, which is usually caused by the heart beating too fast or irregularly, but the heart rate could be higher than normal when the person is jogging. Accordingly, auxiliary sensors set in a playground also provide some support data for the judgment. Then, based on this combination, a risk function could be designed. However, it is important to note that risk functions are quite different under different application scenarios. For example, a totally different risk function should be adopted to detect whether an elderly person is falling asleep at the end of a transfusion. Accordingly, in this article, we do not give a specific definition of risk function, but use the idea of quantification.

The output of risk function is a threshold, denoted by risk factor. Higher risk factor indicates a higher degree of abnormalcy in the elderly person. Once it is larger than a given value, an alarm is sent to RHCP. This given value is risk threshold, denoted as R_{th}. For example, if the range of risk factor is set as [0, 100], and risk threshold R_{th} is set as 90, the elderly person is considered to be in a normal state when output varies between [0, 90]. When output is larger than 90, however, an alarm is sent to report that the person is in an abnormal state.

In LDPA, risk function will serve as the utility function of RDAA. Data from primary sensors must be processed by risk function to obtain final results. Accordingly, the correctness of risk functions is very important. To ensure correctness, risk functions could initially be designed for some typical and widely used scenarios. Then clinical experiments should be launched with these risk functions, based on which some other features could be added to improve their reliability and accuracy. However, note that the complexity also increases with more completed features, so it is important to select a proper algorithm for data analysis. In this article, we adopt RVNS to randomly select a local optimal solution and reconstruct neighborhood structure to improve analysis efficiency with high accuracy.

Terminal Condition of RVNS — Here we set the terminal condition of RVNS. There are two conventional terminal conditions for RVNS: maximum time of neighborhood reconstruction and maximum CPU running time. If we adopt the former as terminal condition, RVNS will stop once the maximum time is achieved. At this moment, because the latest data still need to be analyzed, a relatively large risk factor obtained earlier and its relevant data are invalid, which increases the number of neighborhood reconstruction. On the other hand, if the maximum CPU running time is adopted as the terminal condition, this problem is still unsolved.

It is notable that, in this article, once risk factor is larger than risk threshold, LDPA sends an alarm to RHCP, and risk factor may reach its peak at this time. However, if there are some risk factors that are larger than risk threshold but smaller than the current highest risk factor, these risk factors are ignored in the following analysis, and relevant abnormal data are ignored without an alarm.

Accordingly, we need to change conventional terminal conditions. We set the terminal condition of RVNS as a given threshold. Even if risk factor is smaller than the former peak, an alarm signal is generated in the next RDAA running turn when it is larger than that given threshold.

RDAA — After level division, RDAA will construct a neighborhood structure according to the current global optimal solution. Specifically, RDAA will randomly select an initial global optimal solution (the current global optimal solution) from the solution space: a column of S, denoted as X. The neighborhood structure is constructed according to the level of X, denoted as $l(opt)$ (opt $\in [1, max]$), and the neighborhood structure constructed according to X is denoted as $N(X)$. If the level of a data group collected at a specific sampling time is denoted as $l(temp)$ ($temp \in [1, max]$), data in this group belongs to $N_k(X)$, where $k = |l(temp) - l(opt)|$.

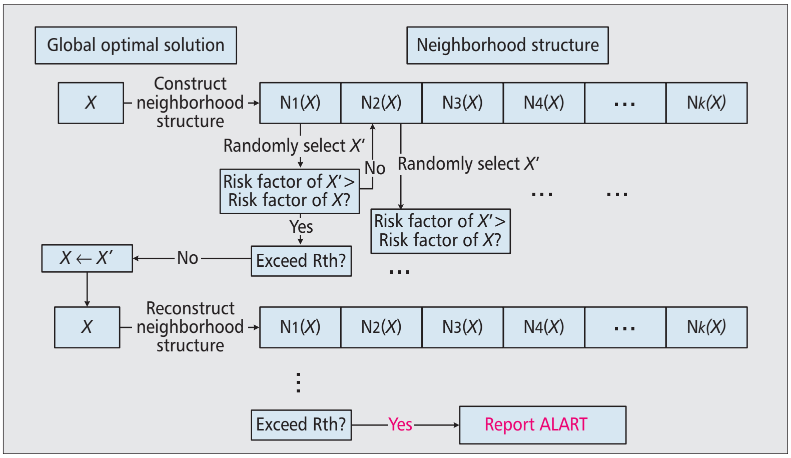


Figure 5. The core process of RDAA.

Before the start of RVNS, every group of data in the buffer is divided into a specific neighborhood according to the initial global optimal solution (current global optimal solution). The initial global optimal solution is processed by risk function, and the risk factor of the initial global optimal solution is recorded. When a new RVNS turn is launched, a local optimal solution is generated; X' is generated from the first neighborhood randomly. If the risk factor of X' is larger than that of X, the relevant data of X' need to be paid more attention. In this case, the current global optimal solution X is replaced by the local optimal solution X', and the neighborhood structure is reconstructed according to a new global optimal solution. Specifically, when risk factor is larger than risk threshold, an alarm is sent to RHCP. On the other hand, if the risk factor of the local optimal solution X' is smaller than the current global optimal solution, RVNS continues to search in the next neighborhood. The whole process of RDAA is illustrated in Fig. 5.

Further Design of RDAA — It is noted that abnormality may not be found in one RVNS run due to random search. To solve this problem, a twice-division buffer is adopted. The buffer is divided into several sections; then a section is divided into several slots. In this way, data may have enough time to wait for analysis. Meanwhile, the analyzing time for one run of RVNS is much shorter than that of other ergodic algorithms. Therefore, the timeliness of the proposed algorithm can be ensured even though data need to be analyzed through several RVNS runs. In the end, this section also serves as the releasing buffer unit. By transferring the oldest data into the database, the latest data can be provided enough space.

Essentially, RDAA is an algorithm to solve a queuing problem [14]. In this scenario, the local server is a servant and received data is a client. According to basic theory in the queuing problem, if the arrival rate of a client is larger than the service rate of a servant, the queue tends to

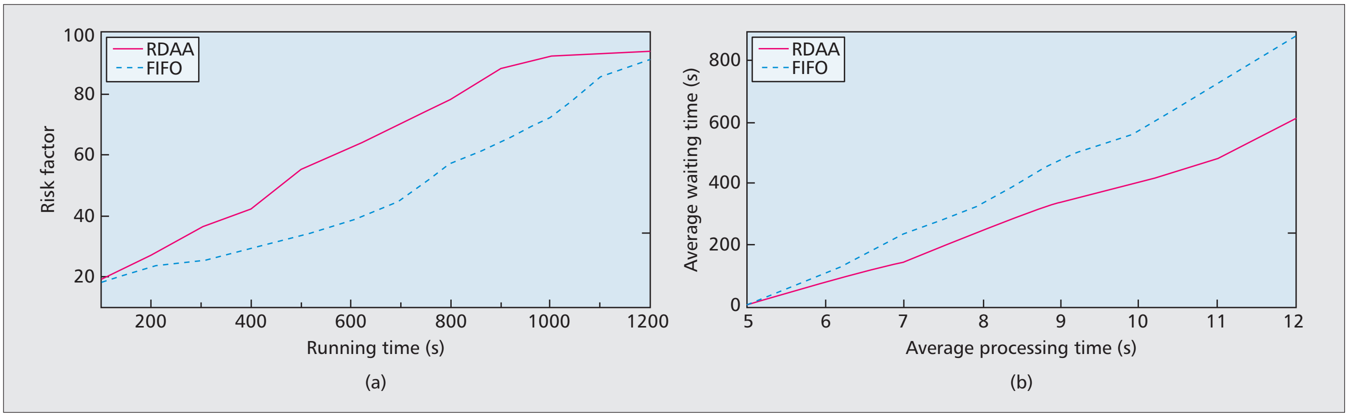


Figure 6. Performance comparison: a) risk factor vs. running time; b) average waiting time vs. average processing time.

infinity. However, because comprehensive monitoring of the elderly will induce explosively increasing data, a traditional data analysis algorithm is inapplicable. Hence, RDAA is proposed by reorganizing the queue and analyzing representative data.

Performance Evaluation

In this section, we compare the performance of RDAA with FIFO [12] through simulation. Data are aligned in chronological order in FIFO. In RDAA, data in several neighborhoods can be dealt with in a short period through reorganizing data into a neighborhood structure. RDAA randomly chooses data to analyze in a systematically changeable neighborhood structure. In this case, there is a higher chance to find abnormal data.

In this simulation, the Arrhythmia Data Set in the UCI Machine Learning Depository is utilized as a data source for heart rate [15], which is also regarded as the source of level division. Risk factor can be obtained by analyzing the heart rate in different scenarios. Therefore, we set two scenarios for elderly behavior: indoor home and outdoor playground. Several sensors are deployed in these two scenarios, which detect different environmental factors such as temperature, humidity, air pressure, lighting, and rainfall. Combining these factors into the scenarios, risk function in RDAA can return a risk factor. The range of risk factor is [1, 100], and risk threshold is set as 90.

In the simulation, we perform a local server through C++, in which a group of data are received every 5 s. We carry out two groups of simulation, the results of which are demonstrated in Figs. 6a and 6b, respectively. In the first group, we compare the highest risk factor returned from the beginning of simulation through RDAA and FIFO. The range of time consumption is from 5 to 10 s, and sampling time is at each 100 s from 0 to 1200 s. In the second group, we observe the average waiting time of data with different average processing time from 5 to 12 s in different algorithms.

Due to the massive amount of data, processing time is longer than the arriving interval of data. In this case, data increase in the traditional FIFO queue, which increases average waiting time and decreases efficiency. In RDAA, however, algorithm performance is improved due to the reorganization of data through neighborhood structure. RDAA can return abnormal data in less time, and average wait time is shortened.

Conclusion

To solve the problem of rapid analysis and processing of massive data in RHCP increasing network load in AAL, we propose LDPA to share the workload of RHCP. Meanwhile, RVNS is introduced to the field of data processing. A proper data analyzing algorithm, RDAA, based on the data from a filter and a classifier, is designed in AAL through changing the conventional terminal conditions of RVNS. In further research, we will pay more attention to the design of risk function under different scenarios, and simplifying risk function on the premise of accuracy.

Acknowledgment

This work has been partially supported by NSFC (61100213), NSF of Jiangsu (BK20141427), SRFDP (20113223120007), PAPD (yx002001), Huawei Innovation Research Program (YB2014010048), Educational Commission of Guangdong Province, China Project (2013KJCX0131), and 2013 Special Fund of Guangdong Higher School Talent Recruitment.

Lei Shu is the corresponding author for this article.

References

[1] A. Hristova and A. M. Bernardos, "Context-Aware Services for Ambient Assisted Living: A Case Study," *Proc. First Int'l. Symp. Appl. Sci. Biomed. Commun. Tech.*, 2008, pp. 1–5.

[2] P. Hansen and N. Mladenovic, "Variable Neighborhood Search: Principles and Applications," *Euro. J. Oper. Res.*, vol. 130, no. 3, 2001, pp. 449–67.

[3] A. Bamis and D. Lymberopoulos, "The BehaviorScope Framework for Enabling Ambient Assisted Living," *Pers. Ubiquitous Comput*ing, vol. 14, no. 6, 2010, pp. 473–87.

[4] A. Munoz and J. C. Augusto, "Design and Evaluation of an Ambient Assisted Living System Based on an Argumentative Multi-Agent System," *Pers. Ubiquitous Computing*, vol. 15, no. 4, 2011, pp. 377–87.

[5] J. McNaull and J. C. Augusto, "Data and Information Quality Issues in Ambient Assisted Living Systems," *J. Data Info. Qual.*, vol. 4, no. 1, 2012, p. 4.

[6] A. McNaull and D. Pietro, "Situation Awareness in Applications of Ambient Assisted Living for Cognitive Impaired People," *Mobile Net. Appl.*, vol. 18, no. 3, 2013, pp. 444–53.

[7] M. A. Hossain and P. K. Atrey, “Modeling and Assessing Quality of Information in Multisensor Multimedia Monitoring Systems,” *ACM Trans. Multimedia Comp. Commun. Appl.*, vol. 7, no. 1, 2011, p. 3.

[8] J. Winkley and P. Jiang, “An Ambient Assisted Living Platform,” *IEEE Trans. Consumer Elect*., vol. 58, no. 2, 2012, pp. 364–73.

[9] A. J. Jara and M. A. Zamora-Izquierdo, “Interconnection Framework for Mhealth and Remote Monitoring Based on the Internet of Things,” *IEEE JSAC*, vol. 31, no. 9, 2013, pp. 47–65.

[10] F. Zhou and J. Jiao, “A Case-Driven Ambient Intelligence System for Elderly In-Home Assistance Applications,” *IEEE Trans. Sys. Man Cybernetics C, Appl. Rev*., vol. 41, no. 2, 2011, pp. 179–89.

[11] M. Mulvenna and W. Carswell, “Visualization of Data for Ambient Assisted Living Services,” *IEEE Commun. Mag*., vol. 49, no. 1, 2011, pp. 110–17.

[12] R. I. Davis and S. Kollmann, “Controller Area Network Schedulability Analysis with FIFO Queues,” *Proc. ECRTS*, 2011, pp. 110–17.

[13] L. Zhou, J. J. Rodrigues, and L. M. Oliveira, “QoE-Driven Power Scheduling in Smart Grid: Architecture, Strategy, and Methodology,” *IEEE Commun. Mag*., vol. 50, no. 5, May 2012, pp. 136–41.

[14] M. L. McManus and M. C. Long, “Queuing Theory Accurately Models the Need for Critical Care Resources,” *Anesthesiol.*, vol. 100, no. 5, 2004, pp. 1271–76.

[15] H. A. Guvenir, B. Acar, and H. Muderrisoglu, “Arrhythmia Data Set in UCI Machine Learning Repository,” http://archive.ics.uci.edu/ml/datasets/Arrhythmia; UC Irvine, School of Information and Computer Science, 1998.

Biographies

Kun Wang [M] is currently a postdoc fellow in tje Electrical Engineering Department, University of California, Los Angeles (UCLA). He received his Ph.D. degree from the School of Computer, Nanjing University of Posts and Telecommunications, China, in 2009. He is also an associate professor in the School of Internet of Things at the same university. He has published over 50 papers in related international conferences and journals, including *IEEE Communications Magazine*, IEEE GLOBECOM 2013, IEEE ICC 2014, and so on. His current research interests include wireless sensor networks, delay-tolerant networks, stream computing, ubiquitous computing, mobile cloud computing, and information security technologies.

Yun Shao is a postgraduate student in information networks at Nanjing University of Posts and Telecommunications. His current research interests include wireless sensor networks, delay-tolerant networks, and stream computing.

Lei Shu [M] received his Ph.D. degree from the Digital Enterprise Research Institute, National University of Ireland, Galway, in 2010. Until March 2012, he was a specially assigned researcher in the Department of Multimedia Engineering, Graduate School of Information Science and Technology, Osaka University, Japan. In October 2012, he joined Guangdong University of Petrochemical Technology, China, as a full professor. in 2013, he started to serve Dalian University of Technology as a Ph.D supervisor. Meanwhile, he is also the vice-director of the Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, China. He is the founder of the Industrial Security and Wireless Sensor Networks Lab. His research interests include wireless sensor networks, multimedia communication, middleware, and security. He received the GLOBECOM 2010 and ICC 2013 Best Paper Awards. He has servd as Editor-in-Chief of *IEEE CommSoft E-letter* and *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems.*

Guangjie Han [M] is currently a professor in the Department of Information and Communication Systems at Hohai University, China. He was a visiting research scholar of Osaka University from October 2010 to October 2011. He finished his work as a postdoctoral researcher in the Department of Computer Science at Chonnam National University, Korea, in February 2008. He worked at ZTE Company from 2004 to 2006, where he held the position of product manager. He received his Ph.D. degree from the Department of Computer Science, Northeastern University, Shenyang, China, in 2004. He has published over 110 papers in related international conferences and journals. He has served on the Editorial Boards of 13 international journals, including the *Journal of Internet Technology* and *KSII Transactions on Internet and Information Systems*. He has served as a Co-Chair for more than 20 international conferences/workshops and a TPC member of more than 50 conferences. He holds 41 patents. He has served as a reviewer of more than 50 journals. His current research interests are sensor networks, computer communications, mobile cloud computing, multimedia communication, and security. He is a member of ACM.

Chunsheng Zhu received his B.E. degree in network engineering from Dalian University of Technology, China, in June 2010 and his M.Sc. degree in computer science from St. Francis Xavier University, Antigonish, Nova Scotia, Canada, in May 2012. He has been working toward his Ph.D. degree in the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada, since September 2012. He has had around 40 papers published or accepted by refereed international journals (e.g., *IEEE Transactions on Industrial Electronics*, *IEEE Systems Journal*) and conferences (e.g., ICC). His current research interests are mainly in the areas of wireless sensor networks and mobile cloud computing.

Asynchronous Flow Scheduling for Green Ambient Assisted Living Communications

Dan Wu, Yuming Cai, and Mohsen Guizani

ABSTRACT

AAL applications that support users (in particular, the elderly and patients) in staying at home are increasingly becoming popular in health care systems. Basically, these AAL applications provide personal information through location-aware, data-aware, and context-aware heterogeneous devices such as sensors and actuators. Obviously, due to this heterogeneity, it is very important to study the asynchronous flow scheduling problem. This article designs a simple but efficient asynchronous flow scheduling scheme aiming to sense, predict, and realize AAL applications. Specifically, the scheduling architecture is developed by analyzing various device characteristics and user activities, and the corresponding applications are classified from user needs aspects. Subsequently, we propose asynchronous flow scheduling taking into account energy efficiency and implementation simplicity. Detailed discussions of the proposed system and possible future research directions are provided.

INTRODUCTION

Ambient assisted living (AAL) communication is an innovative transmission technique derived from the Internet of Things and ubiquitous computing. Essentially, AAL communication constitutes a classical signal communication and processing environment that is active and responsive to user behavior [1]. Recently, substantial technologies on the Internet of Things and ubiquitous computing have been developed for different applications, such as smart city, environmental protection, and e-health. AAL communication, from the functional perspective, can be viewed as a specific e-health technology to increase communication quality of service (QoS) and decrease communication overhead [2].

Importantly, AAL systems can be employed for monitoring and improving the health conditions of patients. For example, medication management and reminders can help patients increase their self-care capacities [3]. In addition to helping patients, AAL is also able to provide sophisticated safety services for the elderly, for example, video surveillance, audio ordering, abnormal behavior detection, and emergent help [4]. Life services are also included in the AAL system (e.g., personal training plans, important event reminders). Moreover, AAL communications can enable users (in particular, the elderly and patients) to stay connected with others via video or audio communications [5].

To provide personalized services according to the communication environment and user needs, AAL communication usually possesses the following characteristics:

- An AAL environment should be a heterogeneous platform that leverages ubiquitous computing, wireless sensor networks, video/audio processing, and information security technologies.
- AAL should be context-aware to perceive user needs or request a dynamic communication environment.
- AAL should be an adaptive system that naturally interacts and changes with users and devices.
- AAL should be energy-efficient to prolong the service life of the system.

Unfortunately, the above characteristics have not been clearly investigated in the context of the AAL environment so far, and they form the main technical challenges for broad applications of AAL communication [6].

In this work, we aim to take a step toward understanding how the above characteristics impact the performance of AAL communication. In particular, we analyze the system architecture and applications, and propose a novel asynchronous delivery mechanism based on the heterogeneous data arrival. Notably we provide a unified theoretical analysis method in the context of AAL, and design a distributed flow scheduling scheme to realize the theoretical analysis. Interestingly, the scheduling scheme is simple enough for online operation. Specifically, we investigate:

- How heterogeneous flows can be scheduled efficiently in the context of asynchronous arrivals
- How to further reduce the energy consumption in the course of asynchronous delivery
- How to apply this distributed flow scheduling algorithm to practical AAL applications

Dan Wu and Yuming Cai are with PLA University of Science and Technology.

Mohsen Guizani is with Qatar University.

0163-6804/15/$25.00 © 2015 IEEE

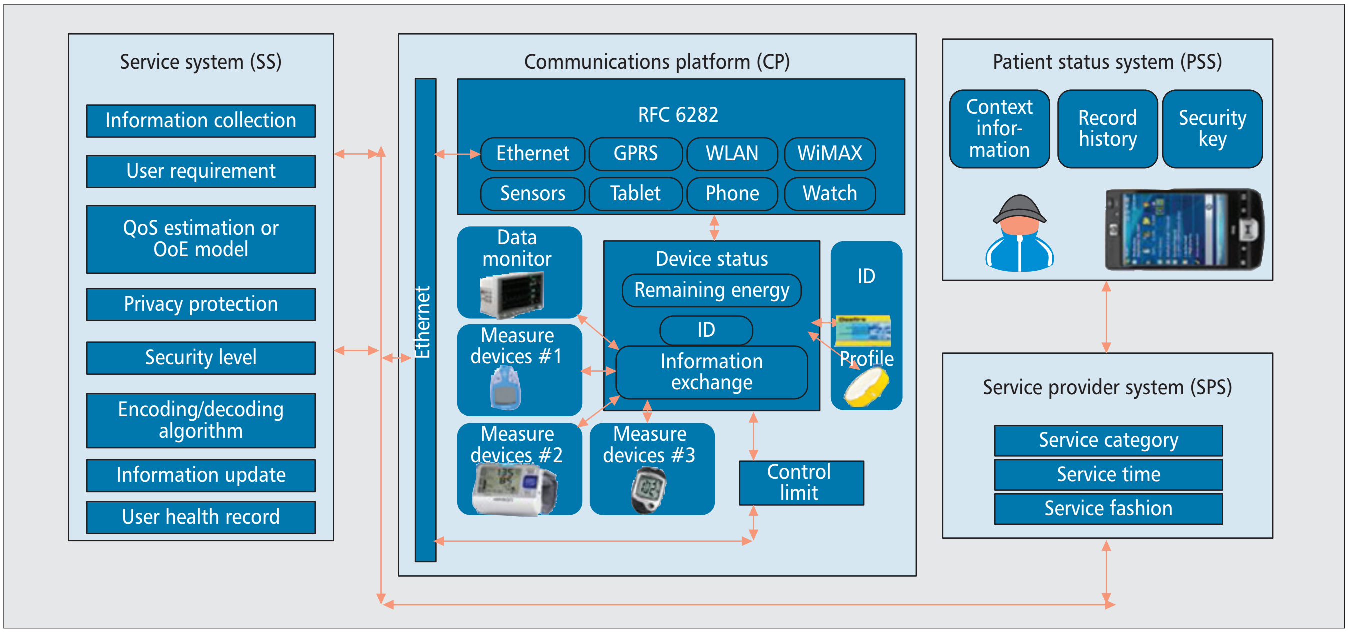


Figure 1. A typical AAL system architecture.

These questions are explicitly investigated and answered in this work.

The rest of the article is organized as follows. The system architecture is introduced in the following section. Then possible AAL applications are introduced. Following that, we propose a novel asynchronous flow scheduling by taking energy efficiency into account. Detailed implementation issues and possible future directions are then discussed. Finally, concluding remarks are provided in the last section.

Architecture

In Fig. 1, we design an AAL system architecture composed of two basic components: communications platform (CP) and service system (SS). Specifically, the CP provides a communication environment to support heterogeneous and asynchronous data collection, and acquire useful data from the AAL environment. As for the SS component, it is composed of a combination of e-health services and data security guarantees for wireless and mobile integration of devices. To reduce the amount of transmitted data, the SS also drops off the packets with the same content from different devices or the same device at different periods. Additionally, from the functional perspective, the AAL system architecture also plays an administrative control role for the fault detector/report, remote control, and join or leave other AAL systems [7].

It should be noted that the proposed AAL system architecture can easily be integrated with existing telemedicine or e-health, especially for rehabilitation training systems, which are roughly divided into two parts: a patient status system (PSS) and a service provider system (SPS). Essentially, the PSS and SPS are very similar, in nature, to the CP and SS in the AAL environment. In the remainder of this section, we describe the functions of the CP and SS, and show how to integrate them with the current PSS and SPS.

Communications Platform

The communication platform is based on the UDP traffic type over IPv6 adapted to RFC 6282 [6]. Specifically, it designs an efficient information expression method (e.g., header compression) for data exchange. In order to enhance the connectivity and reliability of the CP, various devices are employed to provide a ubiquitous environment. Note that here the term ubiquitous does not mean to place a large number of devices in the AAL environment, but represents the use of enough devices to support all the heterogeneous flows.

It is well known that traditional protocols, such as IEEE 1073/HDP, are not appropriate for AAL communication since they typically consume relatively large amounts of energy and yield high overhead. The main problems lie in the following:

- All the collected data should be aggregated at the sink sensor, and this sink sensor is too busy to schedule all the flows satisfying heterogeneous QoS requirements.
- The transmitted data from each device has a large proportion of replication. Therefore, a novel signal-processing-based transmission protocol (SPBTP) based on IEEE 1073 is proposed.

Essentially, SPBTP is based on the correlation between the transmitted data to reduce the overload and optimize the transmission size. This process is recursive, since it is initially assumed that all the data from different devices are different as well. Therefore, the information exchange process leads to a time delay for calculating the similarity of the data. Also, it yields more information overhead at the sink device for comparison. To overcome these shortcomings, SPBTP employs a distributed structure in which it carries out preprocessing at each device to analyze the relevant parts from the crude data to compress the collected data, and makes its

To further lighten the burden of the sink device, SPBTP adopts a linear optimization method to obtain an optimal sample frequency to achieve the trade-off between information accuracy and energy consumption, and this generates security and integrity characteristics for AAL communication.

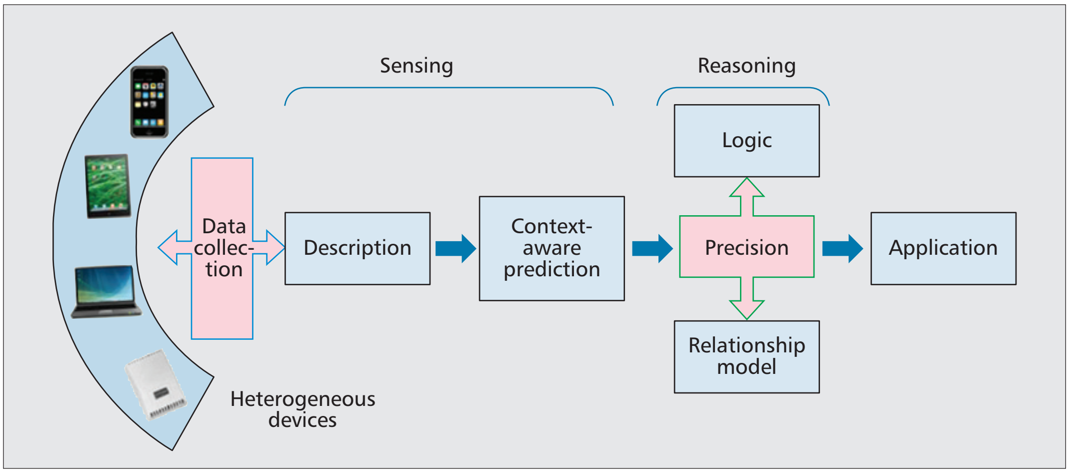


Figure 2. Characteristics of the AAL application.

continuous and real-time transmission feasible. Moreover, to further lighten the burden of the sink device, SPBTP adopts a linear optimization method to obtain an optimal sample frequency to achieve the trade-off between information accuracy and energy consumption, and this generates security and integrity characteristics for AAL communication.

Service System

Analogous to the SS for sensor networks, the core part of AAL communication is context-aware modeling in which various context models, representing diverse communication scenarios, are set up for a formal description of different contextual information. In addition, for some special scenarios, context understanding through a reasoning engine (RE) is a necessary step at the reasoning layer. The RE entails a mapping procedure from the unknown space (new scenarios) to known space (known scenarios). This procedure is so crucial that it determines the overall QoS, and it is normally implemented in three steps; specifically, finding solutions of the most related actions by utilizing an online comparison algorithm, implementing the understanding commands by decomposing the comparison results into known actions, and performing the physical operations if the decomposed action is the same as the known one. It should be noted that the scenario-driven SS is structured as a closed loop, which is adaptive to the dynamic environment, responsive to heterogeneous flows, and consequently suitable for AAL communication.

Moreover, information security is one of the most important factors of the SS. Specifically, it integrates a suitable security stack based on the given CP, which supports heterogeneous flows with heterogeneous devices. As a result, it ensures the security of users' information via the CP. From the perspective of its role, the SS is also relevant to providing a secure content service, because it is fundamental to guarantee no data drops during the communication, and integrate different levels of security in a unified SS.

To this end, we employ a status control strategy (SCS). The SCS offers security features for any communication in the AAL communication environment, ensuring anonymous information protection from external devices and the privacy of patient data [6, 8]. This function is very similar to pervasive mobile health care, since it also provides support for different service requirements from different devices. In this case, the service software can be utilized directly from the health care system.

Applications

Typically, applications in the AAL environment should be appropriate for heterogeneous devices, such as body sensors, smartphones, and actuators. Conceptually, AAL applications acquire and collect data from various devices using different transmission protocols and signal processing methods to diverse users with different QoS requirements (Fig. 2). The main application scene for AAL is e-health, but it does not preclude other applications such as tele-learning and online shopping. In the context of e-health, each user plays a distinct role in the AAL application due to service requirements, transmission method, and user satisfaction. As a result, it is impossible to design a general application, in principle, for any user. Instead, we just provide a typical example to illustrate how the application works in the remainder of this section.

AAL applications, both software and hardware, have evolved from relatively simple emergency call service into sophisticated and smart information process and analysis services that support users with immediate, short-term, and long-term care. With the evolved procedure, the processed data and computational complexity have increased tremendously. Thanks to advances in integrated circuits, smart computation, and communications technologies, devices are able to provide the anticipated processing ability accordingly. In terms of designing an appropriate AAL application, it is important to understand or predict users' requirements and practices, which are also referred to as behavior.

It is important to note that there is a need for AAL applications to visually demonstrate collected information in an intuitional manner to all the users while satisfying the information

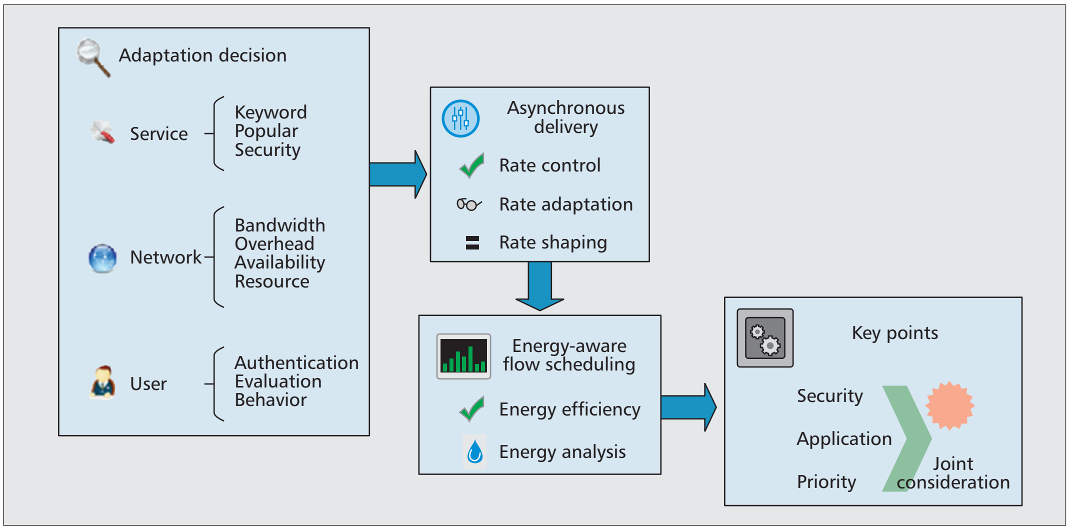


Figure 3. Framework of asynchronous flow scheduling.

It is important to note that there is a need for AAL applications to visually demonstrate collected information in an intuitional manner to all the users while satisfying the information security requirements. Inevitably, this manner heavily relies on techniques of signal processing and visualization.

security requirements [9]. Inevitably, this manner heavily relies on techniques of signal processing and visualization. For example, normal users who implement AAL applications are likely to have little knowledge of computers or medicine. Generally speaking, data collected in AAL applications can include various information to realize the following functions:

- Provide an alert service for an emergency event (e.g., a heart attack
- Offer an alert service for deviant actions and behaviors (e.g., falling down)
- Remind users of their physical training or mental relaxation

Moreover, AAL applications should learn and update users' behavior all the time. Accurate prediction of user behavior can provide rich information in designing an efficient data transmission protocol [10]. Note that the transmission strategy proposed in [10] plays a very important role in the AAL environment, and can be viewed as the basis for data transmission jointly considering energy and spectrum efficiencies. For example, with the help of predicting user behavior, it is not difficult to design an energy-efficient data transmission method in which repeated data (i.e., when a user falls asleep) cannot be transmitted at that time.

Currently, there are several software packages providing AAL applications that include some parts of the mentioned elements [8]. For example, the most popular software visually represents data from different devices and shows information with various importance levels. While this system can offer urgent information alerts once any abnormal event occurs, it cannot provide intelligent analysis for the reason of the event. Actually, it is a complex reasoning process that depends on the techniques of machine leaning and online judgment. Obviously, the introduction of these new features will yield substantial computation and energy overhead, which is an unexpected obstacle for the application of AAL. To resolve this problem, it is imperative to design an efficient motion prediction algorithm, online or offline, to reduce the computational complexity of AAL applications. In the next section, we propose a novel scheduling scheme to realize this objective.

Asynchronous Flow Scheduling

In this section, we design an asynchronous flow scheduling scheme by jointly considering the asynchronous flow delivery mechanism and energy-aware flow scheduling as shown in Fig. 3. In particular, an asynchronous flow delivery mechanism mainly deals with heterogeneous flows from heterogeneous devices, and energy-aware flow scheduling reduces energy consumption by introducing advanced signal processing technologies.

Asynchronous Delivery Mechanism

The asynchronous flow delivery mechanism (AFDM) makes use of IEEE 802.11 in order to collect data from each device and the location information of each device relative to the sink node. The asynchronous delivery mechanism bases its adaptation decision on the time of the received data, data correlation, packet loss rate, and security level. The delivery mechanism, to adapt to the nature of the AAL environment, should be distributed, and consists of flow-side and user-side parts. In terms of the flow side, different data from various devices are decoded at different quality levels (QLs) as well. As the devices may move away from the sink point, the received data becomes asynchronous. To resolve this difficulty, the communication area of a device is divided into multiple areas according to the arrival time of the data. Subsequently, each area has a corresponding QL, which denotes the maximum received quality when the entire flow can be received given the transmission delay. From the feedback of users, each device can determine the relative position of the sink device and the arrival time difference of each flow (note that the flow in this work represents data from different devices). Of course, the feedback information also contains the indication of the packet loss. Once receiving no feedback in two or three consecutive time slots, the packet loss can be detected easily. As such, the QL can be

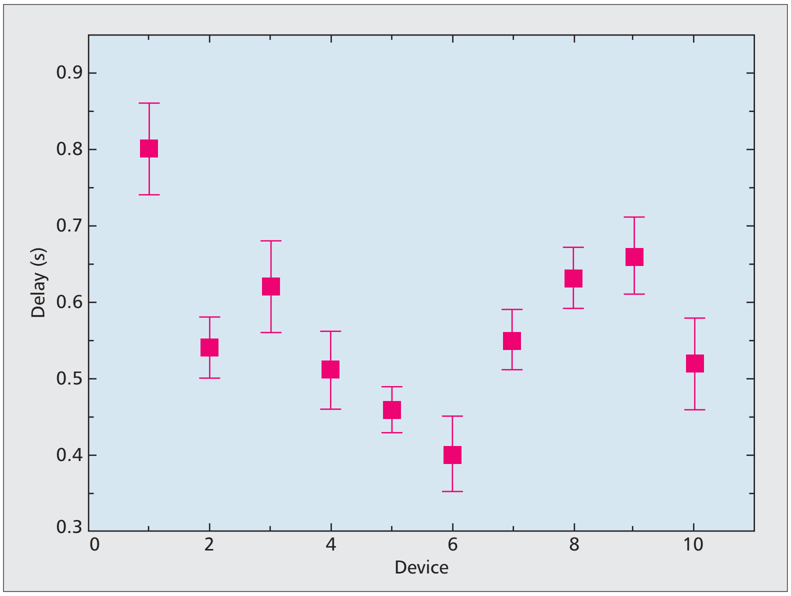


Figure 4. Data arrival delay for each device.

decreased by one level accordingly, which guarantees that the maximum QoS can be achieved by each user. Similarly, QL can be increased once consecutive feedback can be detected. Intuitively, AFDM increases the data transmission quality by adaptively setting the transmission strategy according to different arrival times and positions.

Data adaptation aims to resolve the issues that impact the QoS of the application, and it is driven by communication environment conditions (packet loss rate, maximum transmission rate, etc.), device status conditions (e.g., remaining power, transmission distance), and user behavior (preference, habits, etc.). Historically, a large number of adaptive flow solutions have been designed by taking into consideration one or some of the above mentioned issues. Unfortunately, it is worth noting that they only address some specific problems, and no general solutions have been proposed yet for a universal AAL communication environment.

To realize the asynchronous delivery mechanism, current data delivery solutions addressing wireless network-related issues adapt the transmission rate or throughput of the link to the available bandwidth or power. Roughly speaking, these methods can be divided into three categories: rate control, rate adaptation, and rate shaping [8]. But no matter which method, they have the following common shortcomings:

- They do not possess constant performance when the flows are asynchronous.
- They usually cost nontrivial communication overhead and computational complexity, which are unacceptable for a general AAL communication scenario.

To solve the above technical challenges, in this work we subtly introduce the flow waiting strategy [10] into AFDM. Specifically, we first set the priority level of each flow according to the content importance. If a flow with high priority arrives first at the sink device, it is scheduled immediately; when a flow with low priority arrives, it waits for the high-priority flow to be processed. Essentially, this strategy is so simple that it yields almost no additional overhead, and it can achieve constant performance at the end user.

Energy-Aware Flow Scheduling

Since heterogeneous devices play an extremely important role in the AAL system, energy-aware flow scheduling is a necessary part of system operation. Recently, substantial energy-aware flow scheduling algorithms have been proposed in the context of sensor networks. However, these algorithms cannot be applied directly to the AAL environment due to the following reasons:

- The security level of the flow scheduling in the AAL is much higher. As we know, a higher security level usually employs more complex encryption algorithms, which inevitably lead to higher energy consumption. Although the quantitative relationship between the security level and energy consumption is not yet clear, the qualitative relationship is certain.
- The application scenario of the flow scheduling in the AAL is much more complicated. Usually, heterogonous devices compose the AAL environment; hence, the device relationship between them is complex as well. For example, sensors of body temperature and heartbeat are strongly related to each other. As a result, how to take advantage of this relationship to reduce energy consumption is an interesting topic.
- The priority level of the flow scheduling in the AAL is much more flexible. The AAL is a typical application-driven system, and applications normally vary in user preference to service level. In this case, the priority level of flows is dynamic, which also yields higher energy consumption for the sensors. Therefore, how to efficiently predict flow priority is beneficial for energy efficiency.

In this part, we propose a hybrid energy-aware flow scheduling algorithm by jointly considering the security level, application scenario, and priority level. Specifically, we first set an appropriate security level for a specific service. For operational simplicity, we predefine a table of various encryption algorithms for different levels of security. As such, the energy consumption of the encryption algorithm can be precisely designed in the AAL environment. Then we sort the heterogeneous devices into multiple categories according to the application scenario (e.g., video sensors, audio sensors, data sensors). In particular, the energy consumption of each sensor category can be estimated approximately; for example, the energy consumption of the video sensor is usually larger than that of the audio sensor. Subsequently, the priority level of the flow is set by the application, scenario, and user. Different priority levels correspond to different energy consumption levels (usually the higher the priority, the higher the importance), which also means it can consume more energy.

With this information at hand, we can easily apply them to the traditional energy-aware scheduling algorithms. Here, we emphasize three design criteria in the AAL environment:

•The definition of the security level should be based on practical applications and user needs. In particular, data privacy should be fully considered by introducing a specific user access mechanism. For example, patients' data can only be accessed by their specific doctors or themselves. Other people cannot have the right to know and utilize such data.

•To accommodate the flexible application scenario, a flow estimation strategy should be involved to reduce the energy consumption due to scenario change. This case is particularly important for dynamic scenarios such as people running and playing games.

•Similar to the security level, the priority level should be designed according to user preferences and specific applications. In order to avoid subjective errors, it is necessary to introduce the concept of quality of experience (QoE) and provide a reference score. As such, a systematic subjective test should be conducted via at least 100 users, and different kinds of applications and scenarios should be considered. Then these results can be averaged to obtain a relatively objective priority level.

To verify the efficiency of the proposed scheduling scheme, we employ the simulation settings in [7]. Specifically, five kinds of flows are employed, and their priorities are set according to [8]. There are 10 devices in the AAL communication system. For simplicity, each device has the same transmission capacity. As the benchmark, we compare our scheduling with other two flow scheduling algorithms, media-aware scheduling [8] and energy-aware scheduling [10]. Figure 4 plots the data arrival delay for each device, and Fig. 5 shows the improvement of energy efficiency. These results clearly demonstrate the efficiency of the proposed scheme.

Implementation Issues

Implementing the flow scheduling algorithm for the AAL system requires special attention, because of the characteristics of the user and the physical limitations of the devices. Conceptually, the flow scheduling mechanism should deal with the trade-off relationship between the users and devices. On one hand, users' feedback should be automatically adapted for the scheduling decision. On the other hand, not all the feedback should be considered due to the existence of abnormal users. In this sense, user behavior detection should be recorded and updated in the whole scheduling process. To accomplish this task, online effort technology [5] that only requires small samples of user feedback can be involved in the scheduling mechanism. A specific normal behavior model can be established via an online learning algorithm; as a result, the user interface is kept as simple and friendly as possible. For example, the possibility of irrational users can be rejected, flow congestion can be avoided by limiting user feedback, and parallel computing can be operated via the distributed data collection.

Moreover, sleep time of the device plays an important role in energy-aware flow scheduling in AAL. Generally, long sleep time (data sampling is low) reduces the precision of the device measurements, while short sleep time (data sampling is high) leads to high energy consumption. Therefore, a uniform sleep time scheme is necessary for scheduling the measurements of the sensors, and it should consider the following factors:

- The remaining energy of each device
- The flow state of each device
- Future possible flow congestions
- How the possible flow congestion impacts the energy consumption

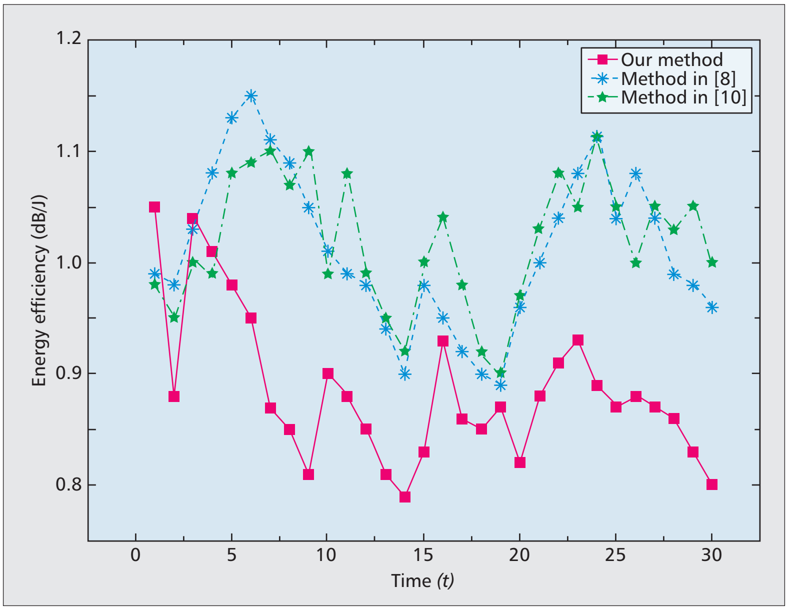


Figure 5. Performance comparison in terms of energy efficiency.

To deal with these factors, the modeling of the device sleep time is desired. In particular, device relative position and movement direction can be modeled as a Markov process [10]. Because devices in the AAL communication usually have a certain computational capacity, one possible solution is to enhance the communication capacity with the help of the computing capacity. For example, according to the sleep model, the devices can predict possible flows via the amount of computation, and as a result, an efficient scheduling can be realized.

Discussions

To extend the application and improve the QoS of AAL communication, in this section, we discuss several important issues that can be embedded in future AAL communication. It should be noted that possible implementation of these technologies is still a challenge. The most challenging aspect, to the best of our knowledge, is how to simplify/modify them to adapt to the context of the AAL communication platform.

Data security and user privacy: From the above discussion, it is obvious that AAL communication brings new concerns about data security and user privacy. In this case, future AAL communication should employ a variety of data protection, user authentication, and data verification methods based on signal, biometric, and physiological characteristics. Meanwhile, lightweight data encryption algorithms can also be embedded into AAL communication, and a scalable security level of AAL communication can be designed to obtain a reliable and secure communication link.

Advanced signal processing method: Network coding, as an advanced coding method that takes into consideration the network state, has been successfully applied to general wireless networks. But how to utilize this technology in an AAL environment to reduce energy consumption is still an open problem. Moreover, compressed sensing technology can also be used to reduce the sampling frequency and transmitted data.

Sophisticated robotics technology: Until now, a large number of intelligent robots do not support health care work due to the high cost problem. In the future, on one hand, the cost of robotics will be reduced substantially due to the development of hardware chips. On the other hand, more users' education should be conducted in terms of the acceptance of robots in helping the elderly and patients. It is clear that future robots will not only be capable of helping users to handle their daily matters, but also able to learn actions and behaviors of users and make intelligent decisions on their behalf.

Corresponding standards and regulations: So far, there is no international communication standard for AAL. Therefore, we expect that it will take a long time to pave the way to adopt AAL communications for mass production and use. Also, there are no structured regulations for AAL implementations or misconduct of complex AAL systems. To protect user rights, the user should be clearly informed of the possible consequences of AAL communication. This means that the current AAL communication technology cannot guarantee a high successful communication ratio. In addition, to reduce users' misconduct, it is necessary to provide sufficient training in order to ensure safe and efficient system usage.

Current AAL communication technology can not guarantee a high percentage of a successful communication ratio. In addition, to reduce users' misconduct, it is necessary to provide sufficient training in order to ensure safe and efficient system usage.

Conclusion

This article aims to define and evaluate important metrics to characterize the QoS in the context of AAL communications. To this end, we propose a simple but efficient asynchronous flow scheduling scheme aiming to sense, predict, and realize AAL application. Specifically, the scheduling architecture is developed by analyzing various device characteristics and user activities, and the corresponding applications are classified from the aspect of the user needs. Subsequently, we propose an asynchronous flow scheduling by taking into account energy efficiency and simple implementation. Moreover, potential applications are analyzed and possible solutions provided. Finally, some challenges are identified for future investigation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 61301163, 61301162), and Jiangsu Provincial Natural Science Foundation of China (Grant BK20130067).

References

[1] P. Rashidi and A. Mihailidis, "A Survey on Ambient-Assisted Living Tools for Older Adults," *IEEE J. Biomed. Health Informatics*, vol. 17, no. 3, May 2013, pp. 579–90.

[2] M. Mulvenna *et al*., "Visualization of Data for Ambient Assisted Living Services," *IEEE Commun. Mag.*, vol. 49, no. 1, Jan. 2011, pp. 110–17.

[3] F. Zhou *et al*., "A Case-Driven Ambient Intelligence System for Elderly in-Home Assistance Applications," *IEEE Trans. Syst. Man, Cybern. C, Appl. Rev*., vol. 41, no. 2, Mar. 2011, pp. 179–89.

[4] P. Gyorke and B. Pataki, "Energy-Aware Measurement Scheduling in WSNs Used in AAL Applications," *IEEE Trans. Instrum. Meas*., vol. 62, no. 5, May 2013, pp. 1318–25.

[5] P. Barsocchi *et al.*, "Evaluating Ambient Assisted Living Solutions: The Localization Competition," *IEEE Pervasive Computing*, vol. 12, no. 4, 2013, pp. 72–79.

[6] A. Coronato and G. De Pietro, "Tools for the Rapid Prototyping of Provably Correct Ambient Intelligence Applications," *IEEE Trans. Software Eng.*, vol. 38, no. 4, 2012, pp. 975–91.

[7] O. R. E. Pereira, J. Caldeira, and J. Rodrigues, "Body Sensor Network Mobile Solutions for Biofeedback Monitoring," *Mobile Networks and Applications*, vol. 16, no. 6, Dec. 2011, pp. 713–32.

[8] L. Zhou et al., "Distributed Scheduling Scheme for Video Streaming over Multi-Channel Multi-Radio Multi-Hop Wireless Networks," *IEEE JSAC*, vol. 28, no. 3, 2010, pp. 409–19.

[9] J. Rodrigues, O. Pereira, and P. Neves, "Biofeedback Data Visualization for Body Sensor Networks", *J. Network and Computer Applications*, vol. 34, no. 1, Jan. 2011, pp. 151–58.

[10] L. Zhou *et al*., "Energy-Spectrum Efficiency Trade-Off for Video Streaming over Mobile Ad Hoc Networks," *IEEE JSAC*, vol. 31, no. 5, May 2013, pp. 981–91.

Biographies

DAN WU received her B.S., M.S., and Ph.D. degrees from the Institute of Communications Engineering, PLA University of Science and Technology, Nanjing, China, in 2006, 2009, and 2012, respectively. She is now a postdoctoral researcher at the same institute. Her research interests are mainly in resource allocation and management, game theory, cooperative communications, and wireless sensor networks.

YUMING CAI received his B.S. degree in physics from Xiamen University, China, in 1982, his M.S. degree in micro-electronics engineering, and his Ph.D. degree in communications and information systems, both from Southeast University, Nanjing, China, in 1988 and 1996, respectively. His current research interest includes MIMO systems, OFDM systems, signal processing in communications, cooperative communications, and wireless sensor networks.

MOHSEN GUIZANI [F] is currently a professor and associate vice president for graduate studies at Qatar University. He was chair of the Computer Science Department at Western Michigan University from 2002 to 2006, and chair of the Computer Science Department at the University of West Florida from 1999 to 2002. He also served in academic positions at the University of Missouri-Kansas City, University of Colorado-Boulder, Syracuse University, and Kuwait University. He received his B.S. (with distinction) and M.S. degrees in electrical engineering, and M.S. and Ph.D. degrees in computer engineering in 1984, 1986, 1987, and 1990, respectively, from Syracuse University, New York. His research interests include computer networks, wireless communications, mobile computing, and optical networking. He currently serves on the Editorial Boards of six technical journals, and the Founder and Editor-in-Chief of Wiley's *Wireless Communications and Mobile Computing* (http://www.interscience.wiley.com/jpages/1530-8669/). He is also the Founder and Steering Committee Chair of the Annual International Conference of Wireless Communications and Mobile Computing. He is the author of seven books and more than 270 publications in refereed journals and conferences. He has guest edited a number of special issues in IEEE journals and magazines. He has served as member, chair, and general chair of a number of conferences. He served as Chair of the IEEE Communications Society Wireless Technical Committee and Chair of TAOS Technical Committee. He was an IEEE Computer Society Distinguished Lecturer from 2003 to 2005. He is a Senior Member of ACM.

Authentication Protocol for an Ambient Assisted Living System

Debiao He and Sherali Zeadally

Debiao He is with Wuhan University.

Sherali Zeadally is with University of Kentucky.

ABSTRACT

Recent advances in healthcare technologies along with improved medical care have led to a steady increase in life expectancy over the past few decades. As a result, we have been witnessing a significant growth in the number of elderly people around the world. Ensuring a comfortable living environment for elderly people has gained much attention in recent years. By leveraging information and communication technologies, the AAL system shows great promise in satisfying many requirements of elderly people and enables them to live safely, securely, healthily, and independently. Over the last few years various AAL systems, mostly based on Wireless Body Area Network technologies, have been proposed to improve the quality of life of elderly people. Since the information transmitted in AAL systems is very personal, the security and privacy of such data are becoming important issues that must be dealt with. We first discuss the overall system architecture of a typical AAL system and its associated security requirements. Next we propose an efficient authentication protocol for the AAL system and describe how it meets various security requirements. Finally we compare the performance of the proposed authentication protocol with two other recent authentication protocols and demonstrate its superior efficiency.

INTRODUCTION

Recent technological developments in farming methods have led to a substantial increase in food production around the world. Today, people can get enough food in most countries. Economic developments in many countries have led to improved healthcare services. Today people can get immediate treatment and help when they are ill. The implementation of various educational initiatives increases people's awareness in preventing diseases and promotes healthy living. As a result of these aforementioned factors, people are now living longer than previous generations. According to a recent United Nations report [1], the number of elderly people in 1950 was 205 million (eight percent of the world's population); by 2013 the number of elderly people had increased to 841 million (11.7 percent of the world's population). The report predicts that the number of elderly people will increase to two billion (22 percent of the world's population) by 2050. With increasing age, many abilities and functions (such as mental activity, physical activity, vision, hearing) of people gradually decrease. Besides, about 80 percent of elderly people older than 65 suffer from at least one chronic disease. Because of these natural health changes, many elderly people cannot take care of themselves. However, a recent survey demonstrates that about 89 percent of elderly people would like to stay in their own house and live independently [2]. To improve the quality of life of elderly people, we need to develop innovative technologies and systems that can help and assist elderly people to take care of themselves. To achieve this goal the concept of Ambient Intelligence (AmI) was recently proposed and has been attracting much attention from researchers and designers. In an AmI system, information and communication technology is used to create networks of different services that could help people in many activities [3]. One specific application of AmI is the Ambient Assisted Living (AAL) system, which can provide healthcare monitoring and tele-health services by leveraging information and communication technologies. Therefore, the AAL system is suitable for helping elderly people to live in their preferred environment independently. To make the AAL system more suitable for practical deployments, many technologies such as Wireless Body Area Networks (WBANs), assistive robotics, e-textile, and mobile and wearable sensors have been used in it [4].

Information exchanged within an AAL system needs to satisfy confidentiality, integrity, and availability requirements. To satisfy these requirements, we need secure communication along with privacy support from the underlying AAL system. The authentication protocol for the AAL system allows only authorized users access to AAL services and ensures secure communication in the AAL system.

We first describe the architecture of a WBAN-based AAL system and the security requirements of the authentication protocol for the AAL system. Next we present a brief review of related authentication protocols and highlight their weaknesses. Then we describe the proposed authentication protocol for the AAL system to address the weaknesses of recent authentication protocols. Finally we present a security and performance analysis of the proposed authentication protocol.

0163-6804/15/$25.00 © 2015 IEEE

Since all information in the AAL system is transmitted on open channels, the authentication protocol used is susceptible to attacks. In this work, our goal is to design a secure, robust and efficient authentication protocol which could resist various types of attacks.

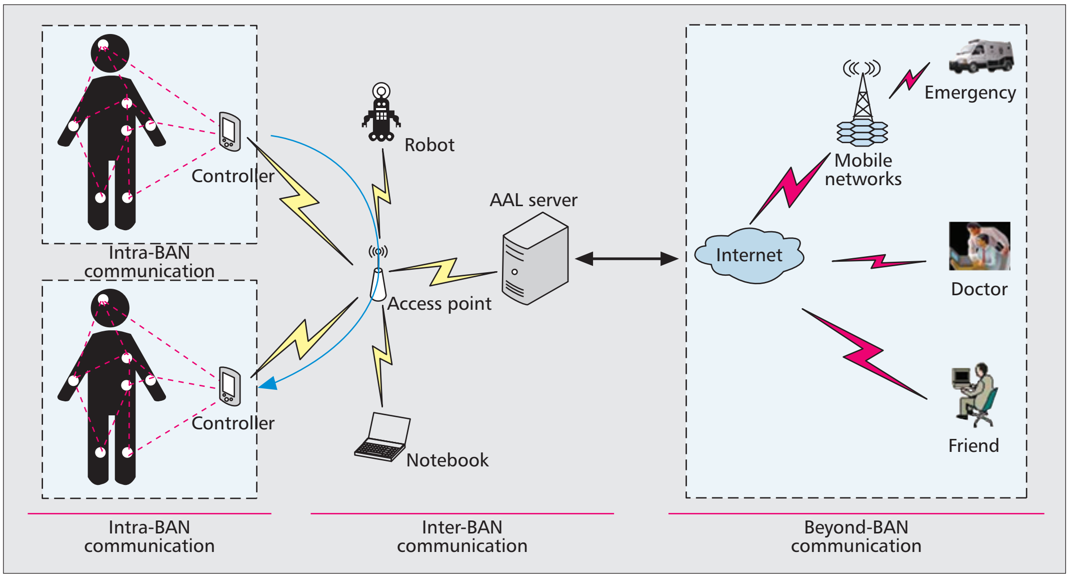


Figure 1. Steps used during the registration process.

System Architecture and Security Requirements of the Ambient Assisted Living System

System Architecture

The concept of the WBAN was first proposed by Zimmerman [5]. By using the wireless personal area network for communication, the WBAN could collect personal biomedical information for many applications through wireless sensor nodes around the body [6]. The IEEE802.15 standard specifies short distance communication standards for the WBAN. The communication system of an AAL system based on WBAN can be organized into three different tiers (see Fig. 1): Intra-BAN, Inter-BAN, and beyond-BAN communications.

Intra-BAN communications denote communications among wireless body sensors and controllers of the WBAN. For example, a sensor monitoring the blood sugar level can communicate with the controller, which is responsible for collecting data of elderly people from sensors attached to them and sending it to remote users. To achieve efficient and secure communication in this tier, the IEEE 802.15.6 standard [7] defines three security levels:

- Level 0 — unsecured communication.
- Level 1 — authentication only.
- Level 2 — both authentication and encryption.

These security levels provide users with adequate security in this tier. Inter-BAN communications denote communications between the controller and devices in elderly people's homes such as notebooks, home service robots, and so on. For example, the controller could communicate with the home service robot through the access point. The beyond-BAN tier connects the AAL server to the Internet. In beyond-BAN communications, users connected to the Internet through wired or wireless connections that can remotely access the AAL server through which they can then access the controllers of the WBAN via the access point. For example, a medical doctor can communicate with the controller through the Internet and the access point.

Security Requirements

Since all information in the AAL system is transmitted on open channels, the authentication protocol used is susceptible to attacks. In this work, our goal is to design a secure, robust, and efficient authentication protocol that can resist various types of attacks. Based on previous works [7–10] about authentication protocols, we identify the following key requirements for a secure authentication protocol for the AAL system.

Mutual Authentication: In order to allow only authorized users to access data collected in the AAL system, mutual authentication among the controller, the AAL server, and the user is needed. This can be achieved with the help of the AAL server.

Anonymity: If an adversary could get the user's identity, the user's privacy will be violated and will cause inconvenience to the elderly people. Thus, anonymity is a key requirement for the authentication protocol for the AAL system.

Non-Traceability: If an adversary could correlate the communication activities of a particular user, he/she could guess the user's real identity with a high probability, thereby violating the user's privacy. Thus, non-traceability is also a key requirement for the authentication protocol for the AAL system.

No Verification (Password) Table: In several previously proposed authentication protocols, the server or the access point keeps a verification (password) table for the authentication process. However, the system can be compromised if the adversary could steal the verification (password) table. Thus, a robust authentication protocol for the AAL system should avoid keeping the password table for authentication purposes.

Session Key Agreement: After mutual authentication, the data transmitted to the controller and the users should be encrypted using the shared session key between them. A shared session key must therefore be generated during the mutual authentication process. Thus, a robust authentication protocol for the AAL system should provide session key agreement.

Perfect Forward Secrecy: Perfect forward secrecy means that an adversary cannot access the session keys generated in previous sessions even if he/she could access the private keys of the user, the controller, and the server. To ensure the secure transmission of information, a robust authentication protocol for the AAL system should provide strong forward secrecy.

Attack Resistance: The authentication protocol is susceptible to various attacks. To ensure secure communication, the authentication protocol should be able to withstand various attacks such as the reply attack, the impersonation attack, the server-spoofing attack, the man-in-the-middle attack, and the modification attack.

Related Work

Security and privacy are important design considerations that must be taken into consideration during the AAL system design. However, authentication research for AAL systems is still in its infancy. To the best of our knowledge no authentication protocol for the AAL system has been proposed. We identified a few authentication protocols designed for the WBAN environment that could be applied to the AAL system after some modifications. Yeh *et al.* [8] proposed the first authentication protocol for the WBAN. In their protocol, the traditional Public Key Cryptography (PKC) is used to achieve their security goal. However, Yeh *et al.*'s authentication protocol is not suitable for mobile devices because of their limited computation and storage capabilities, the PKC used involves the modular exponentiation operation use of modular exponentiation, which may need more computing power and storage space than mobile devices can provide. To improve performance, Liu *et al.* [9] proposed a certificateless anonymous authentication protocol for the WBAN using Elliptic Curve Cryptography (ECC). ECC is suitable for environments with limited computation and communication capabilities because the key size used is smaller. For instance, for a 160 bit key, ECC has the same security level as the 1024bit key with RSA [7]. However, a serious drawback with Liu *et al.*'s protocol is that the access point has to maintain a verifier table for mutual authentication. In this case, the adversary could impersonate the user to get services from the access point by modifying values in the verifier table. Besides, the performance of Liu *et al.*'s protocol is still not satisfactory because of the complex bilinear pairing operation that is used. According to theoretical analysis [10] and past experimental results [11], the computational complexity of a bilinear pairing operation is at least 10 times higher than that of an elliptic curve scale multiplication. Therefore, the performance of Liu *et al.*'s protocol is not satisfactory either. To avoid bilinear pairing operation, Zhao [12] proposed an efficient authentication protocol without using bilinear pairing operation for the WBAN. Zhao's protocol addressed the major weakness in Liu *et al.*'s protocol because no verifier table is used in that protocol. Instead, Zhao's protocol uses pseudo identity instead of the real identity to provide user anonymity. However, Zhao's protocol cannot provide privacy because an adversary could trace the user based on the constant value of the pseudo identity. These previously proposed approaches demonstrate that they cannot satisfy all the security requirements we have identified in the previous section. In contrast, in this work we propose an authentication protocol that can satisfy all the security requirements listed earlier.

Authentication Protocol for an AAL System

As we mentioned earlier, to overcome the weaknesses in previous authentication protocols, we propose an authentication protocol for the AAL system. To avoid using the complex public key infrastructure as in the traditional PKC and verifier table as in the certificateless PKC approach, we use identity-based PKC in the proposed authentication protocol. To avoid using bilinear pairings as in traditional identity-based PKC, we modify the process of generating the user's private key.

There are three participants in the proposed authentication protocol: a controller for the WBAN, the AAL server, and an end-user. First, the AAL server generates the AAL system parameters (such as a finite field, an elliptic curve defined on the field, a generator of the elliptic curve, and a secure hash function) and for ECC, selects a random number as its private key and uses it and the system parameters to compute its public key by computing an elliptic curve scale multiplication of the random number and the selected generator of the elliptic curve. The AAL server derives the controller's private key by computing the hash value of its private key and the controller's identity, where the controller's identity is denoted by its unique name or serial number. Upon receiving the private key, the controller stores it into its memory.

The AAL server derives the controller's private key by computing the hash value of its private key and the controller's identity, where the controller's identity is denoted by its unique name or serial number. Upon receiving the private key, the controller stores it into its memory.

When a user wants to access the data collected by the WBAN in the AAL system, the user must first register with the AAL server by sending his/her identity to it, where the user's identity denotes his/her name. Upon receiving the registration request from the user, the AAL server derives the user's private key by computing the hash value of the user's identity and the AAL server's private key. Then the AAL server sends its public key, the system parameters, and the user's private key to the user through a secure channel, where the secure channel could be established using the Secure Sockets Layer (SSL) protocol. The user stores the AAL server's public key, the system parameters, and his/her private key in the memory of his/her device for future mutual authentication when he/she receives them. The steps involved during the registration process are shown in Fig. 2. It is worth noting that the steps for generating the AAL server's private key and public key are executed only once.

After registering with the AAL server, the user and the controller of the WBAN could authenticate each other with the help of the AAL server. The user is then allowed to access the data collected by the WBAN if he/she is successful with

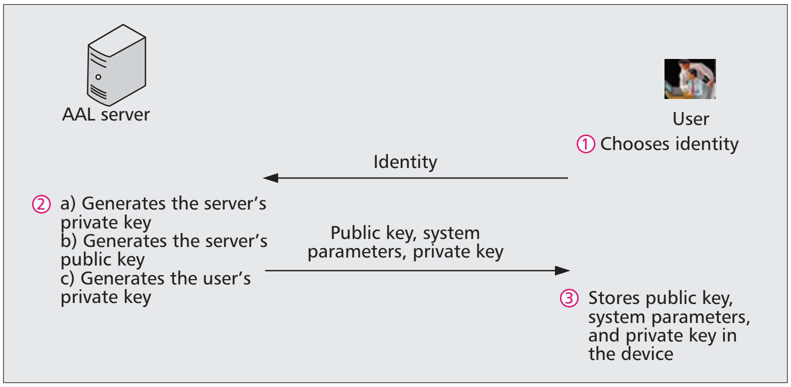

AAL server
User
① Chooses identity
Identity
② a) Generates the server's private key
b) Generates the server's public key
c) Generates the user's private key
Public key, system parameters, private key
③ Stores public key, system parameters, and private key in the device

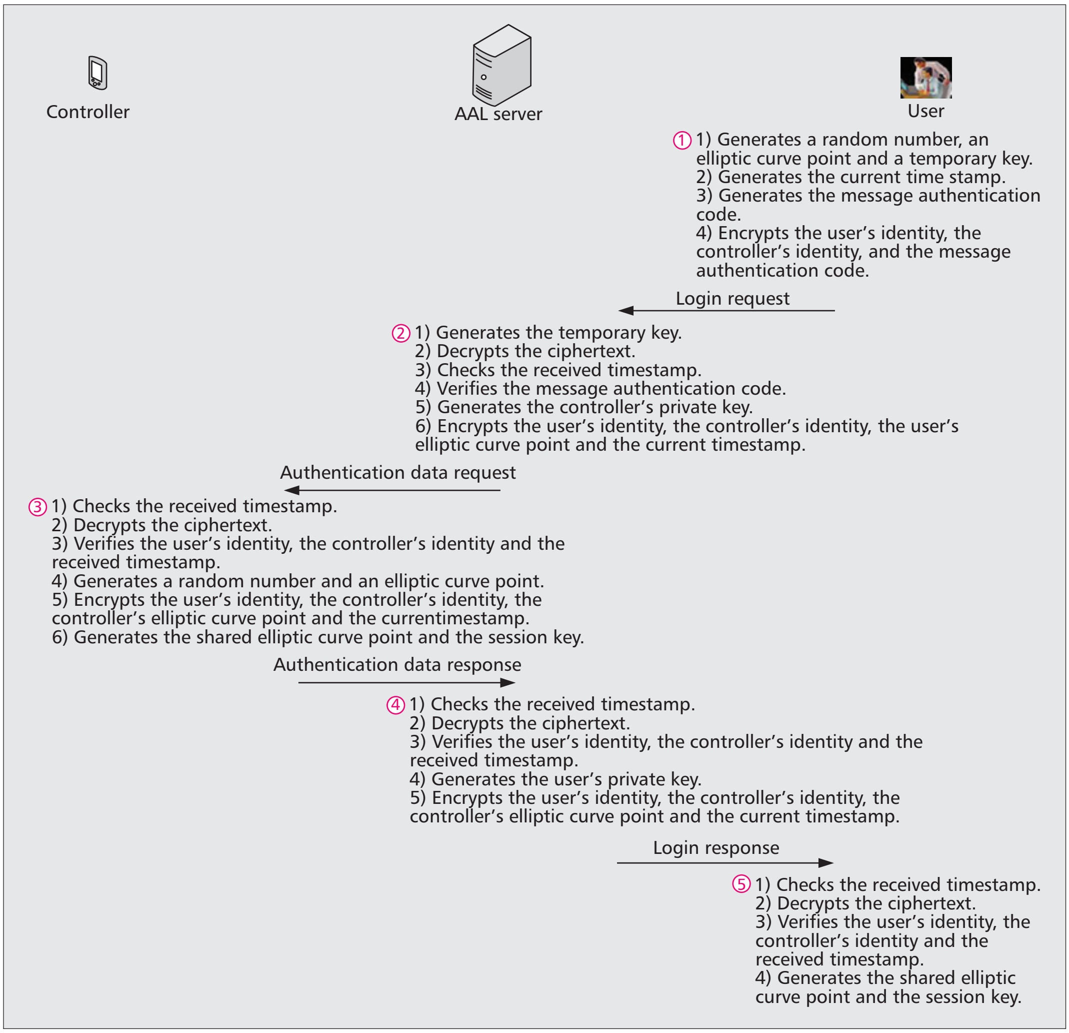
Figure 2. Steps used during the registration process.

the controller's authentication. To withstand the reply attack, the timestamp is used in the proposed protocol. We require that the user and the controller be synchronized with the AAL server. The various steps involved during the process of authentication are shown in Fig. 3. First, the user selects a random number and generates an elliptic curve point on the elliptic curve by computing an elliptic curve scale multiplication of the random number and the generator of the elliptic curve. With the AAL server's public key and the system's parameters, the user generates the temporary key between the user and the AAL server by computing an elliptic curve scale multiplication of the random number and the AAL server's public key. The user generates the current timestamp and uses his/her private key to generate the message authentication code of the user's identity, the controller's identity, the random point, and the selected timestamp. With the temporary key the user uses a symmetric encryption algorithm (such as the Advanced Encryption Standard (AES) algorithm) to encrypt the user's identity, the controller's identity, the selected timestamp, and the message authentication code (such as the keyed-Hash Message Authentication Code (HMAC)). Finally, the user sends the ciphertext and the elliptic curve point to the AAL server as part of the login request.

Upon receiving the login request message, the AAL server performs the following steps to verify the validity of the request. First, the AAL server generates the temporary key between it and the user by computing an elliptic curve scale multiplication of its private key and the received elliptic curve point. The AAL server retrieves the user's identity, the controller's identity, the user's timestamp, and the user's authentication message code by using the temporary key to decrypt the ciphertext. The AAL server checks the freshness of the received timestamp by checking whether the absolute difference between the received timestamp and the current timestamp is larger than some pre-defined value (based on the network speed and reliability). If the timestamp is not fresh, the AAL server rejects the request; otherwise, the AAL server uses its private key and the user's identity to compute the user's private key. Then the AAL server uses the user's private key to verify the validity of the message authentication code. If it is not valid, the AAL server rejects the request; otherwise, the AAL server uses its private key and the controller's identity to compute the controller's private key. With the controller's private key, the AAL server uses a symmetric encryption algorithm to encrypt the user's identity, the controller's identity, the user's random point, and the current timestamp. Then the AAL server sends the cipthertext and the current timestamp to the controller as the authentication data request.

Upon receiving the authentication data request, the controller performs the following steps to verify the validity of the request. The controller first checks the freshness of the received timestamp by checking whether the absolute difference between the received timestamp and the current timestamp is larger than some pre-defined value. If the timestamp is not fresh, the controller rejects the request; otherwise, the controller extracts the user's identity, the controller's identity, the user's elliptic curve point, and the AAL server's timestamp by using its private key to decrypt the received ciphertext. The controller checks if the received timestamp and the one obtained by decrypting the ciphtertext are the same. If they are not the same, the controller terminates the session; otherwise, the controller verifies the validity of the user's identity and the controller's identity. If either of them is not valid, the controller rejects the request; otherwise, the controller accepts the request. To generate the authentication data response, the controller generates a random number, and computes an elliptic curve point by computing an elliptic curve scale multiplication of the random number and the generator of the elliptic curve. The controller uses its private key and a symmetric encryption algorithm to encrypt the user's identity, the controller's identity, its generated random point, and the current timestamp. The controller generates a shared elliptic curve point between it and the user by computing an elliptic curve scale multiplication of its random number and the user's elliptic curve point. Then the controller computes the session key by computing the hash value of the shared elliptic curve point. Finally, the controller sends the cipthertext and the current timestamp to the AAL server as the authentication data response.

Upon receiving the authentication data response, the AAL server executes the following steps to verify the validity of the response. The AAL server first checks the freshness of the received timestamp by checking if the absolute difference between the received timestamp and the current timestamp is larger than some pre-defined value. If the received timestamp is not fresh, the AAL server rejects the response; otherwise, the AAL server retrieves the user's identity, the controller's identity, the controller's elliptic curve point, and the controller's timestamp by using the controller's private key to decrypt the received ciphertext. The AAL server checks whether the received timestamp and the one obtained by decrypting the ciphtertext are the same. If they are not the same, the AAL server terminates the session; otherwise, the AAL server verifies the validity of the user's identity and the controller's identity. If either of them is not valid, the AAL server rejects the response; otherwise, with the temporary key between it and the



① 1) Generates a random number, an elliptic curve point and a temporary key.
2) Generates the current time stamp.
3) Generates the message authentication code.
4) Encrypts the user's identity, the controller's identity, and the message authentication code.

② 1) Generates the temporary key.
2) Decrypts the ciphertext.
3) Checks the received timestamp.
4) Verifies the message authentication code.
5) Generates the controller's private key.
6) Encrypts the user's identity, the controller's identity, the user's elliptic curve point and the current timestamp.

③ 1) Checks the received timestamp.
2) Decrypts the ciphertext.
3) Verifies the user's identity, the controller's identity and the received timestamp.
4) Generates a random number and an elliptic curve point.
5) Encrypts the user's identity, the controller's identity, the controller's elliptic curve point and the currentimestamp.
6) Generates the shared elliptic curve point and the session key.

④ 1) Checks the received timestamp.
2) Decrypts the ciphertext.
3) Verifies the user's identity, the controller's identity and the received timestamp.
4) Generates the user's private key.
5) Encrypts the user's identity, the controller's identity, the controller's elliptic curve point and the current timestamp.

⑤ 1) Checks the received timestamp.
2) Decrypts the ciphertext.
3) Verifies the user's identity, the controller's identity and the received timestamp.
4) Generates the shared elliptic curve point and the session key.

Figure 3. Steps used in the mutual authentication process between the user and the controller.

In the proposed protocol, the AAL server only needs to maintain its private key and does not maintain a verification table for authentication purposes. Therefore, the proposed protocol does not suffer from the drawbacks of hosting a verification table as in previously proposed approaches.

user, the AAL server uses a symmetric encryption algorithm to encrypt the user's identity, the controller's identity, the controller's elliptic curve point, and the current timestamp. The AAL server then sends the cipthertext and the current timestamp to the user as the login response.

Upon receiving the login response, the user performs the following steps to verify the validity of the response. The user first checks the freshness of the received timestamp by checking whether the absolute difference between the received timestamp and the current timestamp is larger than some pre-defined value. If the received timestamp is not fresh, the user terminates the session; otherwise, the user uses his/her temporary key to decrypt the ciphertext and extracts the user's identity, the controller's identity, the controller's elliptic curve point, and the AAL server's timestamp. The user checks whether the received timestamp and the one obtained by decrypting the ciphtertext are equal. If they are not equal, the user terminates the session; otherwise, the user verifies the validity of the user's identity and the controller's identity. If either of them is not valid, the controller rejects the request; otherwise, the user confirms that mutual authentication is achieved. The user also generates a shared elliptic curve point between himself/herself and the controller by computing an elliptic curve scale multiplication of its random number and the controller's elliptic curve point. Then the user computes the session key by computing the hash value of the shared elliptic curve point.

Security Analysis of Proposed Authentication Protocol

In this section we analyze the security of the proposed authentication protocol for the AAL system. We explain how the proposed protocol satisfies all the security requirements described earlier.

Mutual Authentication: Since the selected algorithm for generating message authentication code is a secure algorithm, then it possible to generate the correct message authentication code only if one has the user's private key. Thus, the AAL server could authenticate the user by checking the correctness of the message authentication code when it receives the login request. Besides, because of the selected symmetric encryption algorithm, it is possible to generate the authentication data request/the authentication data response/the login response only if one has the AAL server's private key/the controller's private key/the AAL server's private key. Thus,

> *Even if an adversary could get both the user's elliptic curve point and the controller's elliptic curve point, the adversary cannot generate the shared elliptic curve point between them because of the computational Diffie-Hellman problem. Therefore, the proposed protocol provides strong forward secrecy.*

the controller/the AAL server/the user could authenticate the AAL server/the controller/the AAL server by checking the correctness of the received ciphertext. Therefore, the proposed protocol provides mutual authentication among the user, the AAL server, and the controller.

Anonymity: In the proposed protocol the user's identity and the controller's identity are included in the login request, the authentication data request, the authentication data response, and the login response. However, any adversary cannot get those identities because they are encrypted by a secure symmetric encryption algorithm. Therefore, the proposed protocol could provide anonymity.

Non-Traceability: In the proposed protocol all messages are transmitted in the ciphertext format except for the user's elliptic curve point, which is transmitted in the plaintext format. The user generates a new elliptic curve point randomly for each authentication process. Consequently, any adversary cannot determine the communication activities of a particular user. Therefore, the proposed protocol can provide non-traceability.

No Verification (Password) Table Required: In the proposed protocol the AAL server only needs to maintain its private key and does not maintain a verification table for authentication purposes. Therefore, the proposed protocol does not suffer from the drawbacks of hosting a verification table as in previously proposed approaches.

Session Key Agreement: In the proposed protocol the user and the controller generate a session key by computing the hash value of the shared elliptic curve point between them. Therefore, the proposed protocol could provide session key agreement.

Perfect Forward Secrecy: In the proposed protocol the session key between the user and the controller is generated by computing the hash value of the shared elliptic curve point between them. Even if an adversary could get both the user's elliptic curve point and the controller's elliptic curve point, the adversary cannot generate the shared elliptic curve point between them because of the computational Diffie-Hellman problem. Therefore, the proposed protocol provides strong forward secrecy.

Attack Resistance

We discuss how the proposed protocol can withstand various attacks below.

Replay Attack: In the proposed protocol the current timestamp is used in all messages exchanged among the user, the controller, and the AAL server. A replay attack can be detected by checking the freshness of the received timestamp.

Impersonation Attack: If an adversary wants to impersonate the user/the controller to the AAL server, the adversary needs to generate a correct login request/authentication data response. However, the selected message authentication code algorithm/symmetric encryption algorithm is a secure cryptographic algorithm. As a result, the adversary cannot generate a correct login request/authentication data response because he/she does not know the user's private/the controller's private key. Therefore, the proposed protocol could withstand the impersonation attack.

Server-Spoofing Attack: If the adversary wants to impersonate the AAL sever to the user/the controller, he/she has to generate a correct login request/authentication data request. Without the AAL server's private key, the adversary cannot generate the user's private key and the controller's private key. As a result, the attacker cannot generate a correct login request/authentication data request. Therefore, the proposed protocol could withstand the server-spoofing attack.

Man-in-the-Middle Attack: The proposed protocol provides mutual authentication among the user, the AAL server, and the controller. Therefore, the proposed protocol could withstand the man-in-the-middle attack.

Modification Attack: The user, the AAL server, and the controller check the correctness of the received ciphertext before generating the response to the received message. Since the selected cryptographic algorithms used by the proposed protocol are secure, any modification of the transmitted information will be easily detected. Therefore, the proposed protocol could withstand the modification attack.

Based on the above discussions, the proposed protocol is robust and can mitigate various types of attacks.

Performance Evaluation of the Proposed Authentication Protocol

In this section we analyze the performance of the proposed authentication protocol for the AAL system. As we mentioned previously, we found that several authentication protocols for the WBAN environment could be applied to the AAL system after some modifications. Therefore, we compare the proposed protocol's computational cost in terms of execution time to execute various operations with two of the most recently proposed authentication protocols (i.e. Liu *et al.*'s protocol [9] and Zhao's protocol [12]), which can be applied to the AAL system. For Liu *et al.*'s authentication protocol based on the bilinear pairing, to achieve the same security level as the 1024-bit RSA algorithm, we assume a Tate pairing defined over the supersingular elliptic curve on a 512-bit finite field is used. For Zhao's authentication protocol and the proposed protocol based on the ECC, to achieve the same security level as the1024-bit RSA algorithm, we assume a non-supersingular on a 160-bit finite field is used.

Let T_h, T_{sym}, t_{mm}, T_{exp}, and T_{pair} denote the execution time of one hash function operation, one symmetric encryption or decryption operation, one modular multiplication, one modular exponentiation operation, one elliptic curve scale multiplication, and one bilinear pairing operation, respectively. Using the experimental results obtained in [11, 13], we have the following results: $T_h \approx 0.4T_{mm}$, $T_{sym} \approx 0.4T_{mm}$, $T_{exp} \approx 240T_{mm}$, $T_{ecsm} \approx 29T_{mm}$, and $T_{pair} \approx 620T_{mm}$. In the proposed authentication protocol the user executes one hash function operation, four symmetric encryption or decryption operations, and two elliptic curve scale multiplication operations. Thus, the

	User side	AAL server	Controller	Total Computational Cost
Liu *et al.*'s protocol [10]	$3T_h + 1T_{sym} + 4T_{ecsm} + 1T_{exp} \approx 357.6_{Tmm}$	—	$3T_h + 1T_{sym} + 1T_{ecsm} + 1T_{pair} \approx 650.6T_{mm}$	$6T_h + 2T_{sym} + 4T_{ecsm} + 1T_{exp} + 1T_{pair} \approx 1013.6T_{mm}$
Zhao's protocol [13]	$4T_h + 1T_{sym} + 3T_{ecsm} \approx 89T_{mm}$	—	$5T_h + 1T_{sym} + 6T_{ecsm} \approx 176.4T_{mm}$	$9T_h + 2T_{sym} + 9T_{ecsm} \approx 265.4T_{mm}$
The proposed protocol	$2T_h + 2T_{sym} + 3T_{ecsm} \approx 88.6T_{mm}$	$1T_h + 4T_{sym} + 1T_{ecsm} \approx 31T_{mm}$	$1T_h + 2T_{sym} + 2T_{ecsm} \approx 59.2T_{mm}$	$4T_h + 8T_{sym} + 6T_{ecsm} \approx 178.8T_{mm}$

Table 1. Computational cost comparisons.

execution time of the user is about $2T_h + 2T_{sym} + 3T_{ecsm} \approx 88.6T_{mm}$. The AAL server in the proposed authentication protocol executes one hash function operation, three symmetric encryption or decryption operations, and one elliptic curve scale multiplication operation. Thus, the execution time of the AAL server is about $1T_h + 4T_{sym} + 1T_{ecsm} \approx 31T_{mm}$. The controller in the proposed authentication protocol executes one hash function operation, two symmetric encryption or decryption operations, and two elliptic cure scale multiplication operations. Thus, the execution time of the controller is about $1T_h + 2T_{sym} + 2T_{ecsm} \approx 59.2T_{mm}$. The computational costs between the proposed authentication protocol and two other recently proposed authentication protocols are shown in Table 1.

The results in Table 1 show that the total computational cost of Liu *et al.*'s protocol, Zhao's protocol, and the proposed protocol is $1013.6T_{mm}$, $265.4T_{mm}$, and $178.8T_{mm}$, respectively. The proposed authentication protocol for the AAL system is 5.7 and 1.5 times better in terms of execution time than Liu *et al.*'s and Zhao *et al.*'s protocols, respectively. In addition, Liu *et al.*'s protocol cannot work without a verification table and Zhao *et al.*'s protocol cannot provide anonymity.

Conclusion

We have proposed a novel authentication protocol for the AAL system based on elliptic curve cryptography. In contrast to recent authentication protocols, the proposed protocol not only supports several important security requirements needed by the AAL system, but can also withstand various types of attacks. In addition, the performance analysis results reveal that the proposed authentication protocol is more efficient than the recently proposed authentication protocols.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable comments that helped us to improve the quality and presentation of this article. This research was supported by the National Natural Science Foundation of China (NSFC) under Grants 61370224, 61472074, and 61402339.

References

[1] UN, "World Population Ageing 2013," 2013, pp. 810; available: http://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2013.pdf (last accessed April 30, 2014).

[2] Center for Disease Control and Prevention, "The State of Aging and Health in America 2007," 2007; available: http://www.cdc.gov/aging/pdf/saha_2007.pdf (last accessed April 30, 2014).

[3] D. Ruyter, "Ambient Assisted-Living Research in the CareLab," *Interactions*, vol. 14, no. 4, 2007, pp. 30–33.

[4] P. Rashidi and A.Mihailidis, "A Survey on Ambient-Assisted Living Tools for Older Adults," *IEEE J. Biomed. Health Informatics*, vol. 17, no. 3, 2013, pp. 579–90.

[5] T. Zimmerman, "Personal Area Networks: Near-Field Intrabody communication," *IBM Systems J.*, vol. 35, no. 3/4, 1996, pp. 609–17.

[6] M. Chen *et al.*, "Body Area Networks: A Survey," *Mobile Networks and Applications*, vol. 16, no. 2, 2011, pp. 171–93.

[7] P. Rashidi and A. Mihailidis, "A Survey on Ambient-Assisted Living Tools for Older Adults," *IEEE J. Biomed. Health Informatics*, vol. 17, no. 3, 2013, pp. 579–90.

[8] C. Yeh, H. Chen, and J. Lo, "An Authentication Protocol for Ubiquitous Health Monitoring Systems," *J. Medical and Biological Engineering*, vol. 33, no. 4, 2013, pp. 415–19.

[9] J. Liu *et al.*, "Certificateless Remote Anonymous Authentication Schemes for Wireless Body Area Networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 2, 2014, pp. 332–42.

[10] L. Chen, Z. Cheng, and N. Smart, "Identity-Based Key Agreement Protocols from Pairings," *Int'l. J. Information Security*, vol. 6, 2007, pp. 213–41.

[11] X. Cao and W. Kou, "A Pairing-Free Identity-Based Authenticated Key Agreement Scheme with Minimal Message Exchanges," *Information Sciences*, vol. 180, 2010, pp. 2895–2903.

[12] Z. Zhao, "An Efficient Anonymous Authentication Scheme for Wireless Body Area Networks Using Elliptic Curve Cryptosystem," *J. Medical Systems*, vol. 38, no. 2, 2014, pp. 1–7.

[13] J. Huang *et al.*, "Robust and Privacy Protection Authentication in Cloud Computing," *Int'l. J. Innovative Computing, Information and Control International*, vol. 9, no. 11, 2013, pp. 4247–61.

Additional Reading

[1] U. Esnaola and T. Smithers, "Whistling to Machines," in Ambient Intelligence in Everyday Life," *Lecture Notes in Computer Science Series*, vol. 3864, 2006, pp. 198–226.

Biographies

Debiao He (hedebiao@163.com) received his Ph.D. in applied mathematics from the School of Mathematics and Statistics, Wuhan University, in 2009. He is currently an associate professor in the State Key Lab of Software Engineering, Wuhan, China. His main research interests include cryptography and information security, in particular cryptographic protocols.

Sherali Zeadally (szeadally@uky.edu) is an associate professor in the College of Communication and Information at the University of Kentucky. He received the Bachelor and Doctorate degrees in computer science from the University of Cambridge, England, and the University of Buckingham, England, respectively. He is a Fellow of the British Computer Society and the Institution of Engineering Technology, England.

Reliable MAC Design for Ambient Assisted Living: Moving the Coordination to the Cloud

Elli Kartsakli, Angelos Antonopoulos, Aris S. Lalos, Stefano Tennina, Marco Di Renzo, Luis Alonso, and Christos Verikoukis

Elli Kartsakli, Angelos Antonopoulos, Aris S. Lalos, and Luis Alonso are with Technical University of Catalonia.

Christos Verikoukis is with the Telecommunications Technological Centre of Catalonia.

Stefano Tennina is with the University of L'Aquila, Italy.

Marco Di Renzo is with Université Paris-Sud.

ABSTRACT

AAL technologies constitute a new paradigm that promises quality of life enhancements in chronic care patients and elderly people. From a communication perspective, they involve heterogeneous deployments of body and ambient sensors in complex multihop topologies. Such networks can significantly benefit from the application of cooperative schemes based on network coding, where random linear combinations of the original data packets are transmitted in order to exploit diversity. Nevertheless, network coordination is sometimes required to obtain the full potential of these schemes, especially in the presence of channel errors, requiring the design of efficient, reliable, and versatile MAC protocols. Motivated by the recent advances in cloud computing, we investigate the possibility of transferring the network coordination to the cloud while maintaining the data exchange and storage at a local data plane. Hence, we design a general framework for the development of cloud-assisted protocols for AAL applications and propose a high-performance and error-resilient MAC scheme with cloud capabilities.

INTRODUCTION

Demographic trends of the last decades reveal an indisputable truth: the world's population is gradually getting older. According to the World Health Organization (WHO) [1], the proportion of the global population aged over 60 years is expected to double from about 11 percent in 2013 to 22 percent by 2050. One of the contributing factors to population aging is the fact that, thanks to advances in modern medicine and sanitation, life expectancy tends to increase. In fact, the average life span globally is projected to rise from 69 years in 2005–2010 to 76 years in 2045–2050 [1].

This increasing longevity is unprecedented, and comes with many challenges for individuals and society. Cardiovascular and respiratory diseases, cancer, diabetes, disabilities, and cognitive impairments are chronic conditions that are more prevalent in older age groups. Significant resources are required for the management of these diseases, including frequent and lengthy hospitalizations, long-term monitoring, and complex interactions with medical professionals. As a result, the quality of life of chronic care and elderly patients can be severely compromised, increasing the strain on their families and care givers and placing considerable economic burden on the healthcare system.

The pervasive use of wireless communication technologies can play a fundamental role in improving healthcare delivery, and ensuring cost-effective and patient-centered disease management and prevention. Recently, ambient assisted living (AAL) technologies have emerged as a new paradigm that employs ambient intelligent tools [2] to provide a smart and enhanced living environment. AAL systems promote safety and autonomy, offer assistance with daily activities, and ensure access to social and medical services, thus supporting independent living and encouraging a healthier lifestyle [3]. Typical AAL environments include smart homes [4] and apartments, geriatric or rehabilitation residences, and even hospital wards dedicated to long-term care.

Even though there is no unified framework for the design of AAL systems, some key components of their architecture can be identified. To begin with, unobtrusive health, activity, and ambient monitoring lies at the heart of a typical AAL solution. Wearable or implantable sensors are employed to measure vital signs (body temperature, brain activity, heart rate, etc.) and activity patterns (e.g., walking, fall detection). Ambient sensors are also deployed to obtain context information on the physical environment (temperature, lighting, etc.) and user location. Apart from sensors, AAL systems may include acting devices, such as medical actuators (e.g., insulin pumps), robotic devices, and domotic systems for home automation. All these components are glued together through middleware platforms [5] that integrate software algorithms and reasoning engines for processing and analyzing the collected data [6], provide user-friendly interfaces [7], and facilitate the development of AAL application services.

0163-6804/15/$25.00 © 2015 IEEE

From a communication perspective, an AAL system usually requires the joint deployment of mobile wireless body area networks (WBANs), formed by sensors deployed in the vicinity of or within the body of moving users, and static ambient wireless sensor networks (WSNs). In this heterogeneous scenario, the low-power short-range wireless technologies specified within the standards IEEE 802.15.6 and IEEE 802.15.4 for WBANs and WSNs, respectively, are usually employed for sensor interconnection [8]. Although these standards define the basic mechanisms for channel access and transmissions, there are still many open challenges in order to ensure energy-efficient and reliable communication.

Given the dense deployment of multiple sensors within a specific area, it is possible to achieve significant enhancements by exploiting diversity through node cooperation. The key idea of cooperative schemes is to encourage nodes to act as relays and forward information overheard by neighboring transmissions. Furthermore, with the introduction of network coding, intermediate nodes can combine and process different received information flows in order to achieve performance improvements even in resource constrained networks such as WBANs [9]. However, the benefits of network coding cannot be fully exploited in the presence of channel errors, and under hostile indoor and body area propagation environments the amount of redundancy required to ensure successful decoding can be prohibitive.

An effective way to drastically reduce redundant retransmissions without compromising performance is by enabling the exchange of information among the relays, allowing them to act in a coordinated manner. Hence, novel medium access control (MAC) protocols must be envisioned to handle data transmissions and relay cooperation in such dynamic and multihop topologies. Nevertheless, as deployments tend to grow in complexity, the network coordination becomes overwhelming due to the high number of nodes and the large amount of collected data. Furthermore, distributed approaches may require the nodes to have increasing processing power and storage capabilities.

Recent advances in cloud computing offer an alternative solution to mitigate these challenges. This new paradigm, which is revolutionizing the way information systems are designed, is based on a shared pool of hardware and software resources that are easily accessible via the Internet and often managed by third-party providers. Current research on cloud-based pervasive healthcare is still in its initial steps [10], but it seems that there are tremendous opportunities to be seized through the seamless integration of cloud technology and WBANs [11]. Typical cloud services include data storage, processing, and hosting of medical application services and interfaces [12, 13].

A less investigated but equally intriguing topic is to employ cloud resources at lower communication layers, to design innovative MAC protocols and routing schemes with increased flexibility and reconfigurability. In one of the few existing works in this context [14], a cloud-assisted MAC protocol has been proposed and implemented for a WLAN deployment. The main idea is to transform the access points into a unified user interface and concentrate MAC layer functions to virtual machines provided by cloud services. This novel concept is rather focused on implementation issues, thus offering a significant practical contribution. Another interesting work in [15] gives some preliminary insights on the use of network coding in complex large-scale networks where multiple relay nodes interacting with each other form a wireless cloud. The presented results show the potential gains of different transmission strategies in a very simple topology. The authors also indicate the need for some level of inter-cloud cooperation in order to ensure the best strategy selection by the relays, which can be a challenging task in large-scale networks. Summarizing, these works show that cloud-based solutions for the MAC layer can yield promising results; however, there is a need to establish solid frameworks for the implementation of these schemes.

In this article, we present a novel cloud-based architecture conceived for AAL environments where medical and ambient sensors are densely deployed. We envision a scenario where part of the sensor infrastructure is connected to the cloud, forming a network of cloud-enabled relays that can significantly enhance the flow of information within the system. An innovative feature of our framework is that we transfer the network coordination to a central controller, which is located at the cloud and communicates with the nodes through a cloud-assisted control plane, whereas we maintain the exchange and storage of data packets to a local data plane.

An innovative feature of our framework is that we transfer the network coordination to a central controller, which is located in the cloud and communicates with the nodes through a cloud-assisted control plane, whereas we maintain the exchange and storage of data packets to a local data plane.

The remainder of this article is organized as follows. Initially, we describe the proposed cloud-assisted framework, defining the different operation planes and the key architecture components. We then present a specific case study on the application of network coding in cooperative relay networks to illustrate the significant potential of the proposed framework. After identifying a major performance weakness of network coding schemes in the presence of link failures, we propose a novel cloud-assisted MAC scheme that employs cloud resources to achieve relay coordination and demonstrate, through simulation-based performance evaluation, the significant performance gains that can be obtained. Then we provide some suggestions for future lines of research, stemming from the application of the proposed framework to a wide range of applications for AAL environments. Finally, we close the article with some general conclusions.

A Cloud-Based Framework for AAL

The proposed technology agnostic cloud-based framework defines two planes of operation, as shown in Fig. 1.

The Cloud-Assisted Control Plane — responsible for the network control and coordination. Control signaling has the form of very short

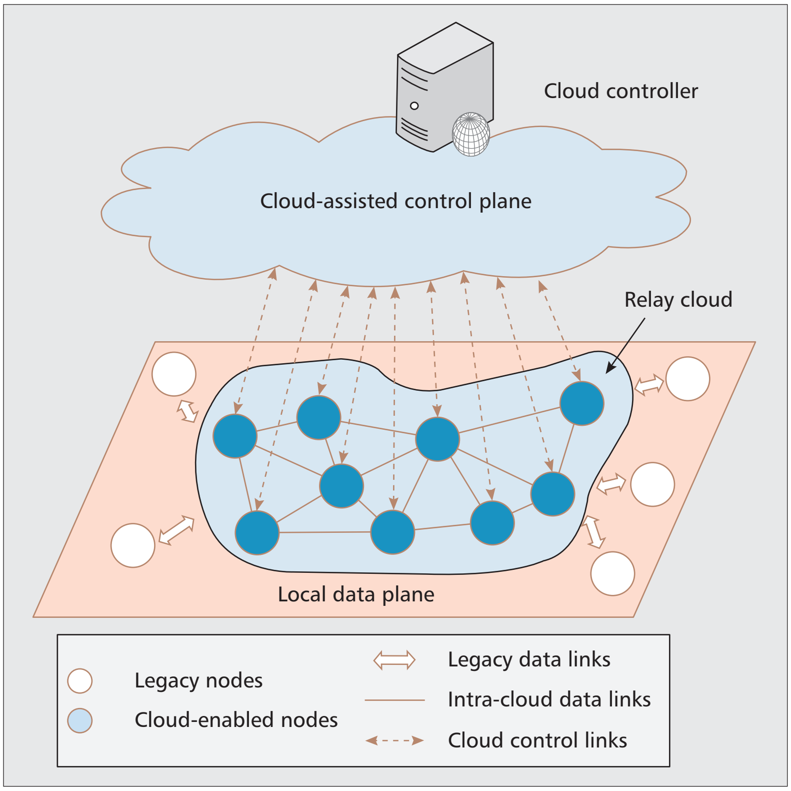


Figure 1. The proposed cloud-based framework.

messages exchanged between the nodes and a coordinating entity, referred to as the cloud controller, located in the cloud. Control communication takes place through a dedicated high-speed link, employing Ethernet or cellular 3G/4G technologies.

The Local Data Plane — responsible for data transmissions among the sensors. To this end, short-range technologies are employed, such as IEEE 802.15.6 for WBANs and IEEE 802.15.4 for WSNs. Without going into detail about the network topology and operation yet, we assume that multihop communication is supported among the sensors in order to achieve a given goal (e.g., data dissemination or data collection by sink nodes). In addition, different protocols can be considered for the MAC layer, including cooperative policies and network coding schemes.

With this system architecture in mind, the network components can be classified into three categories, according to their operation.

The Cloud Controller — which is a coordinating entity located at the cloud, operating only at the control plane. The cloud controller has global knowledge of the network formed by the cloud-enabled nodes and can make optimal decisions for the network operation by processing this information.

The Cloud-Enabled Nodes — which are sensor nodes with enhanced capabilities that can operate in both the data and control planes. These nodes play a key role in the proposed framework. By being able to exchange control information with the cloud controller, they can act in a coordinated manner, forming a cloud of nodes that enables the implementation of advanced cooperative MAC and routing schemes, thus facilitating significantly the flow of information within the data plane.

The Legacy Nodes — which are sensor nodes that can only operate in the data plane. These nodes have more stringent power and processing constraints and can benefit from the enhanced capabilities of the cloud-enabled nodes.

Based on the proposed architecture, the legacy nodes view the network of cloud-enabled relays as a single entity with a unified interface, without the need to be concerned about its structure or individual elements. In addition, crucial global information on the network can be collected by the cloud controller via the control plane. These design features offer several advantages that enable the design of efficient MAC schemes, especially when some level of centralized control is desired.

In the next section, we illustrate one potential application of the proposed framework in AAL environments, by presenting a cloud-assisted MAC scheme that exploits network coding in the presence of channel errors.

Case Study: A Cloud-Assisted MAC Protocol for Enhanced Network Coding Reliability

Recently, the application of practical random linear network coding (RLNC) schemes, based on the generation and transmission of random linear combinations of the original data packets, is showing promising results, offering enhancements in throughput, reliability, and energy efficiency [9]. However, link failures introduced mainly by channel errors can have a detrimental effect on the performance of such schemes. As a solution to this problem, we show how the proposed framework can be applied to the considered scenario and propose CRNC-MAC, an enhanced cloud-assisted MAC protocol that exploits the centralized control capabilities provided by the cloud in order to extract the full potential of RLNC.

Problem Statement: The Need for Coordination in RLNC-Based Schemes

In this case study, we consider an AAL facility where ambient sensors are immersed in the environment and residents are monitored through on-body or implanted medical sensors, forming WBANs. Focusing on a given WBAN, we consider that all the sensed data are gathered by the WBAN coordinator and must, in turn, be forwarded to a locally deployed sink node (e.g., a central medical unit). We also assume that no direct communication is possible between a given source (i.e., the WBAN coordinator) and the final destination; hence, the use of relay nodes is required. In this scenario, we assume that the role of relays is played by the network of ambient sensors, which generally has a mesh multihop topology.

We focus on a typical RLNC scheme in which the source generates N random linear combinations of the N uncoded (original) data packets, by multiplying each packet with a random coefficient drawn by a finite Galois field. Each relay creates new linear combinations of the received packets and forwards them to the destination (or to the next cluster of relays, in the case of multihop topologies). Finally, the destination is able to successfully perform decoding and retrieve the original packets, only after receiving at least N independent linear combinations.

This baseline scheme has an inherent weakness in the presence of channel errors, which is illustrated with the help of the example depicted in Fig. 2a. This example focuses on the mth hop of a generic mesh relay network, that is, on the transmissions that take place between the (m – 1)th and the mth relay clusters. Consider that the original source information is contained within four linearly independent packets (A–D) that have reached the (m – 1)th relay cluster. Due to random errors introduced by the channel, each relay has received only a subset of these four packets, which are then encoded into new RLNC combinations and sent to the next hop. However, assume that in this example the only combination that contains packet D is transmitted by relay R_4 and is not received by any of the relays of the mth cluster. As a result, all the packets generated at the next hop are linear combinations of packets A–C, whereas packet D is absent from all transmissions beyond this point. Ultimately, the missing degree of freedom impedes the decoding process at the destination, causing significant performance degradation.

Given that acknowledgment frames are not employed in multicast transmissions, the most common solution to this problem is the transmission of redundant data copies by the source and relays. However, in networks with dynamic topology and no previous knowledge of the channel conditions, the calculation of the optimum number of redundant retransmissions is not a trivial task. Another approach is to exploit cooperative diversity by introducing a high number of relays. Even though this may be possible in AAL scenarios with dense sensor deployments, involving too many intermediate nodes in each transmission will lead to new problems, such as increased interference and energy consumption.

CRNC-MAC: Introducing Relay Coordination through the Cloud

In order to tackle the aforementioned problem in a decisive way, it is necessary to introduce some level of coordination among the relays. To this end, we propose CRNC-MAC, a novel cooperative RLNC-based protocol with enhanced robustness against errors, built within the proposed cloud-assisted framework. Figure 2b gives a high-level description of the proposed MAC protocol. Going back to the proposed framework architecture, we assume, without loss of generality, that the source and destination are legacy nodes, whereas the ambient sensors are cloud-enabled relays. All data transmissions take place at the local data plane, whereas the cloud-assisted control plane is employed for the exchange of control information between the relays and the cloud controller.

The key concept behind CRNC-MAC is to employ cloud resources for the relay coordination, ensuring that all crucial information from the source is propagated without losses through the multiple hops, thus enabling successful decoding at the destination. In particular, the cloud controller has the task of verifying the reception of all required information on a hop-by-hop basis, requesting retransmissions whenever necessary. In the example of Fig. 2b, once the mth hop transmissions are completed, the relays of the mth cluster forward a list of the received packets to the cloud controller via the control plane. If any crucial information is missing, the relays are notified and request a retransmission from the previous cluster.

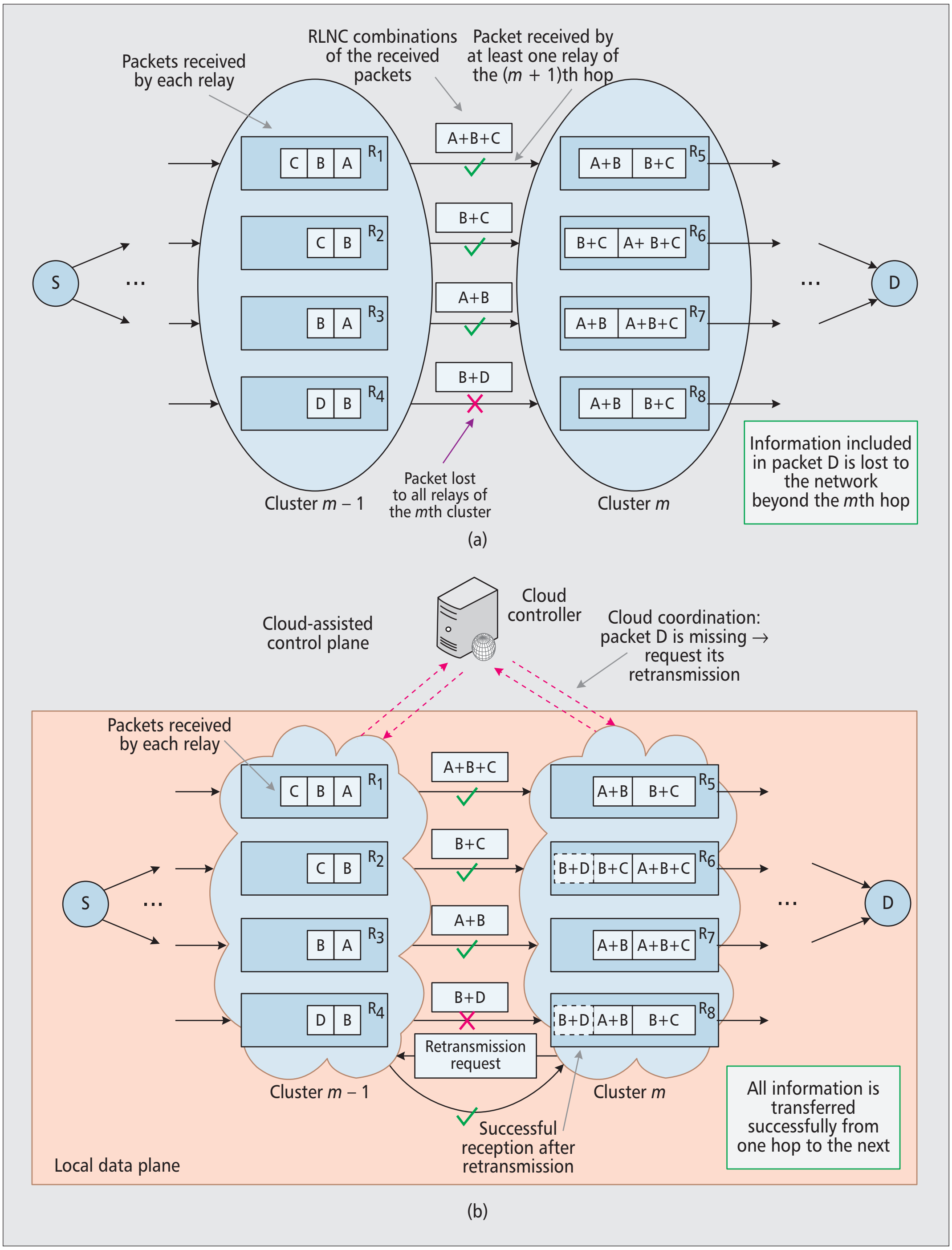
The key concept behind CRNC-MAC is to employ cloud resources for the relay coordination, to ensure that all crucial information from the source is propagated without losses through the multiple hops, thus enabling the successful decoding at the destination.

In continuation, a more detailed description of the protocol operation is given with the help of an example, shown in Fig. 3, where the key steps of the algorithm are numbered and indicated within circles. For the sake of simplicity, we consider a two-hop network in which $N = 4$ original packets are transmitted by the source through the cluster of $R = 3$ cloud-enabled relays.

The protocol operation in each hop is divided into two phases, the *dissemination* and *cloud-assisted coordination* phases. In the transmission phase of the first hop (step 1 in Fig. 3), the source transmits 4 RLNC combinations (packets $P_1 – P_4$) of the original data. The cloud-assisted coordination phase has a variable duration, and involves communication in both data and control planes. In the beginning of this phase, each relay informs the controller of the packets that have been correctly received by transmitting a short control message with the sequence numbers of the respective packets (step 2). Depending on whether all the transmitted packets have been received by the relay cloud, the controller will indicate either the need for retransmissions by the source or the successful termination of the dissemination phase. In our example, packet P_4 has not been received by any of the relays due to channel errors, thus hindering the decoding process at the destination. Hence, through the control plane (step 3), the controller assigns to one of the relays the task of transmitting a request for retransmission (RRT) at the source (step 4). It should be noted that, since the source employs RLNC, it does not need to retransmit the exact missing packets, but only a sufficient number of linearly independent combinations of the original data. In our example, the source transmits one new RLNC packet, P_5 (step 5).

Given that retransmissions are also affected by channel errors, the coordination phase is repeated until all packets are correctly received by the relay cluster. In the considered example, P_5 is received by two of the relays; hence, the dissemination phase is successfully completed. The controller receives an updated packet list (step 6) and issues a transmission schedule for the next hop, dictating the order

We have considered a two-hop relay network in order to gain a more intuitive understanding of the existing performance problems and the potential enhancements of the proposed solution. The two-hop topology also gives us a lower bound of the achievable gain.



Packets received by each relay
RLNC combinations of the received packets
Packet received by at least one relay of the (m + 1)th hop
S
D
R1
C
B
A
R2
R3
R4
A+B+C
B+C
A+B
B+D
R5
R6
R7
R8
A+B
A+ B+C
A+B+C
D
Packet lost to all relays of the mth cluster
Cluster m – 1
Cluster m
Information included in packet D is lost to the network beyond the mth hop
(a)
Cloud-assisted control plane
Cloud controller
Cloud coordination: packet D is missing → request its retransmission
Packets received by each relay
Retransmission request
Successful reception after retransmission
All information is transferred successfully from one hop to the next
Local data plane
(b)

Figure 2. Problem statement and the proposed cloud-assisted MAC solution: a) performance weakness of baseline RLNC schemes under link failures; b) CRNC-MAC provides relay coordination through the cloud-assisted framework.

and the number of packets to be transmitted by each relay (step 7).

In the dissemination phase of the second hop (*step 8*), the relays forward RLNC combinations of their received packets (P'_1–P'_4) according to the schedule. In this case, the destination is a legacy node; hence, there is no cloud-assisted coordination phase. Hence, transmissions take place until the destination receives a sufficient number of copies to decode the original packets and terminates the relaying phase by transmitting a block acknowledgment (BACK).

Performance Evaluation

In this section, we show the benefits of cloud-assisted coordination through a simulation-based performance evaluation of CRNC-MAC. We have considered a two-hop relay network in order to gain a more intuitive understanding of the existing performance problems and the potential enhancements of the proposed solution. The two-hop topology also gives us a lower bound of the achievable gain, since as the number of hops increases, the need for cloud-assisted coordination becomes more imperative. The

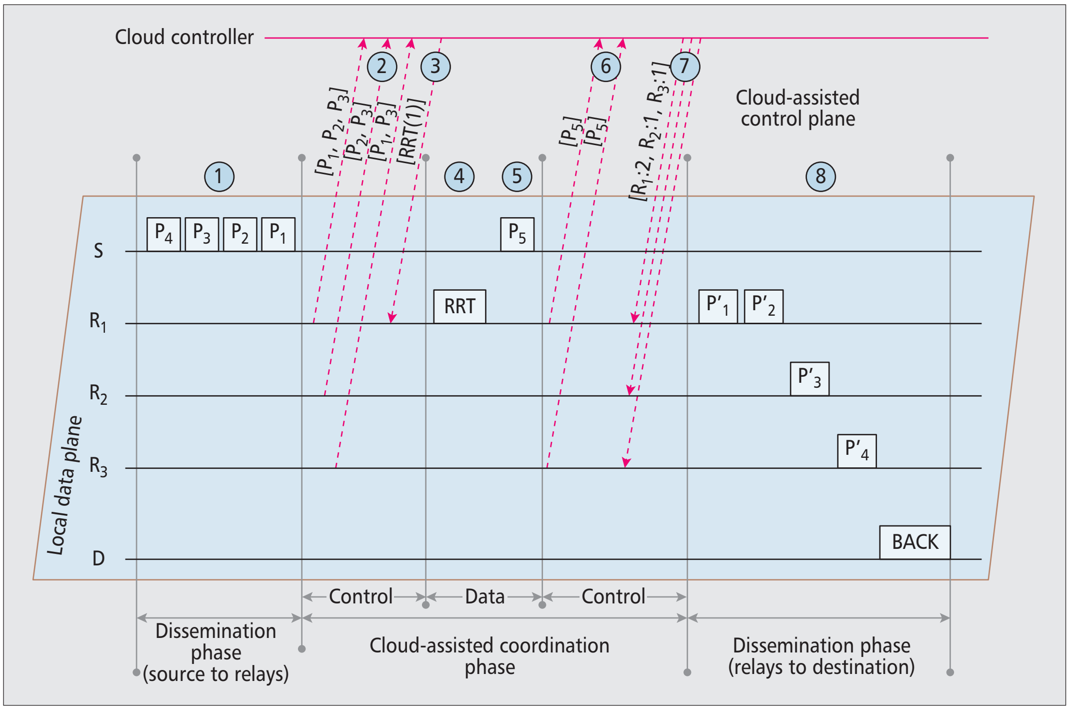


Figure 3. Example of the CRNC-MAC operation.

The key performance gains of CRNC-MAC are appreciated as the error probability grows. ... Successful deconding is always guaranteed, thus yielding an impressive performance gain.

erasure channel is modeled as a Bernoulli process with probability of link failure p. In terms of simplicity, we assume that the relay links are independent, but have similar average channel conditions in both hops, from the source to the relays and from the relays to the destination. Without loss of generality, the narrowband PHY and MAC parameters have been chosen in accordance with the IEEE 802.15.6 specifications for WBANs for data plane communication. The key simulation parameters are summarized in Table 1.

With respect to the control plane, we assume error-free and high-speed communication through Ethernet or cellular third/fourth generation (3G/4G) technologies. In order to quantify the control overhead, we define T_c as the time required for one message exchange between each cloud-enabled relay and the controller. We consider different values for the latency T_c, ranging from zero, for ideal instantaneous control links, up to 10 ms, which is a realistic value for round-trip delays of very short messages.

We compare the performance of CRNC-MAC with a baseline RLNC-based scheme, denoted as RLNC-MAC, which does not support relay coordination. In RLNC-MAC, the relays transmit their encoded data following the IEEE 802.15.6 contention-based access rules. Transmissions are terminated either at the reception of a BACK frame after successful decoding or when the retransmission limit is reached (set to 10 per relay).

The first set of plots in Fig. 4 shows the throughput and energy efficiency performance of CRNC-MAC and RLNC-MAC with respect to the packet error probability p for $R = 3$ relays. Under error-free conditions (i.e., $p = 0$), CRNC-MAC with no latency ($T_c = 0$) achieves an 18 percent throughput improvement with respect to RLNC-MAC, due to the efficient relay transmissions, without collisions and backoff times. However, the key performance gains of CRNC-MAC are appreciated as the error probability grows. In particular, as the channel becomes more hostile, the throughput of RLNC-MAC experiences a steep drop and becomes almost zero for $p = 0.6$. This occurs because even though a percentage of the transmitted packets reaches the destination, there are not enough independent linear combinations to enable successful decoding. On the contrary, CRNC-MAC experiences only a slight drop in performance for high error , due to the increased number of required retransmissions. Nevertheless, successful decoding is always guaranteed, thus yielding an impressive performance gain. Similar results are obtained with respect to energy efficiency, with the achieved gains ranging from 26 percent when no channel errors are considered to more than 3000 percent for $p = 0.6$.

When latency is introduced to the cloud communication, the CRNC-MAC throughput is slightly affected, as shown in Fig. 4a. In particular, with respect to the ideal case when $T_c = 0$, an average throughput degradation of approximately 10 percent is observed when $T_c = 5$ ms, increasing to 20 percent for $T_c = 10$ ms. Hence, a performance trade-off is present, depending on the channel quality and the cloud communication latency. Under very good channel conditions (i.e., with $p < 0.2$), employing cloud resources may not yield significant enhancements, since for high latencies the obtained throughput is practically the same as the baseline RLNC-MAC. Nevertheless, for medium and high error probabilities, the advantages of the cloud-assisted scheme become evident.

The throughput and energy efficiency metrics are also plotted in Fig. 5 as a function of the

Parameters	Values	Parameters	Values
N (packets)	10	RTT, BACK (bytes)	9
L (bytes)	100	Data rate (kb/s)	485.7
Galois field	2^8	Control rate (kb/s)	121.4
PHY preamble (bits)	90	Transmit power (mW)	4.6
PHY header (bits)	31	Receive power (mW)	3.8
MAC header and FCS (bytes)	9	Idle power (mW)	0.712
$[CW_{min},CW_{max}]$	[16, 64]	Sleep power (μW)	4

Table 1. Simulation parameters.

number of relays R for a channel error probability of $p = 0.3$. In this case, CRNC-MAC always outperforms the baseline scheme, regardless of the introduced latency. A very interesting observation is that the throughput performance of CRNC-MAC is not affected much by the number of relays, since the cloud-assisted cooperation enables the achievement of the same diversity gains by a smaller number of relays. As a result, high performance can be achieved even when few relays are available, leading to more energy-efficient communication.

Potential Applications of the Cloud-Assisted Framework

In the previous sections, we have shown how the application of the proposed cloud-assisted framework can enable the design of an efficient MAC scheme that overcomes the performance limitations of RLNC in the presence of channel errors. However, the presented case study is only one example of the benefits that can be obtained through cloud-assisted network coordination. In this section, we identify some potential applications of the proposed framework that can serve as the starting point for further research in the design of enhanced protocols for AAL environments.

Advanced Design for MAC and Network Layer Schemes — enabled by the cloud-assisted centralized coordination and the exchange of control information. Besides the presented case study focused on the application of RLNC schemes, different performance goals can be targeted. Accordingly, the acquired feedback can cover different aspects of the network, such as channel quality, buffer state and content, battery level, node temperature, and so on, opening a vast number of possibilities for the design of enhanced schemes, including channel and energy-aware opportunistic scheduling, collision-free channel access, relay selection schemes, and so on. Some of these schemes can be applied to a wide range of scenarios, whereas others may be specifically designed for healthcare applications. An example of the latter case is temperature-based routing [11], that takes into account the effect of heat dissipation by body nodes in order to minimize tissue heating. This scheme requires efficient clustering, which can be achieved through cloud-assisted coordination.

Flexibility in Network Deployment and Maintenance — given that the internal structure of the relay network is invisible to the legacy nodes. This facilitates the deployment of new cloud-enabled nodes or the performance of maintenance tasks, which may often occur in dynamic AAL environments, and increases robustness against device failures. It should also be noted that there is no need for complex neighbor discovery algorithms to detect any changes in the topology, since this information can readily be provided by the cloud controller.

Mobility and Tracking Support — of legacy nodes in scenarios where the position of the cloud-enabled nodes is known (e.g., fixed deployment or GPS (Global Positioning System) capabilities of the relays). A moving legacy node is likely to interact with different cloud-enabled nodes, thus conveying information about its location. Hence, the cloud-enabled nodes play the role of anchor points. By concentrating all this information at the cloud coordinator, it is possible to recreate the trajectory of the mobile node through localization and tracking algorithms. Such applications often form part of AAL systems for the monitoring of patients moving within a constrained environment.

Distributed Storage Applications — a recent paradigm based on the concept of storing data with redundancy within the network instead of concentrating them at specific sink nodes. Distributed storage is gaining ground as a means to increase reliability in networks with high data contents (e.g., medical records), but requires mechanisms for efficient reconstruction of the original data. The proposed framework can easily support distributed storage applications, where data generated by the legacy nodes are distributed and stored within the relay cloud. In addition, the central coordination can significantly facilitate the data recovery process, since the cloud controller can obtain feedback on the information stored in each cloud-enabled node.

Security — which is an indispensable requirement of healthcare applications, due to the sensitive and confidential nature of medical data. The advantage of the proposed framework is that, by defining two planes of operation for data and control, it can support a wide range of security solutions, either centralized or distributed. Let us consider, for example, two main approaches for key management storage in cloud-enabled WBANs, given in [11], and indicate how their weaknesses are mitigated by the proposed architecture. The first is a centralized approach that employs the cloud resources for the storage of keys, but must rely on the cloud provider to refrain from decoding the encrypted data stored in the cloud. In the proposed framework, this vulnerability risk is not present, since even though the keys are handled by the cloud controller, the data remains on the local data

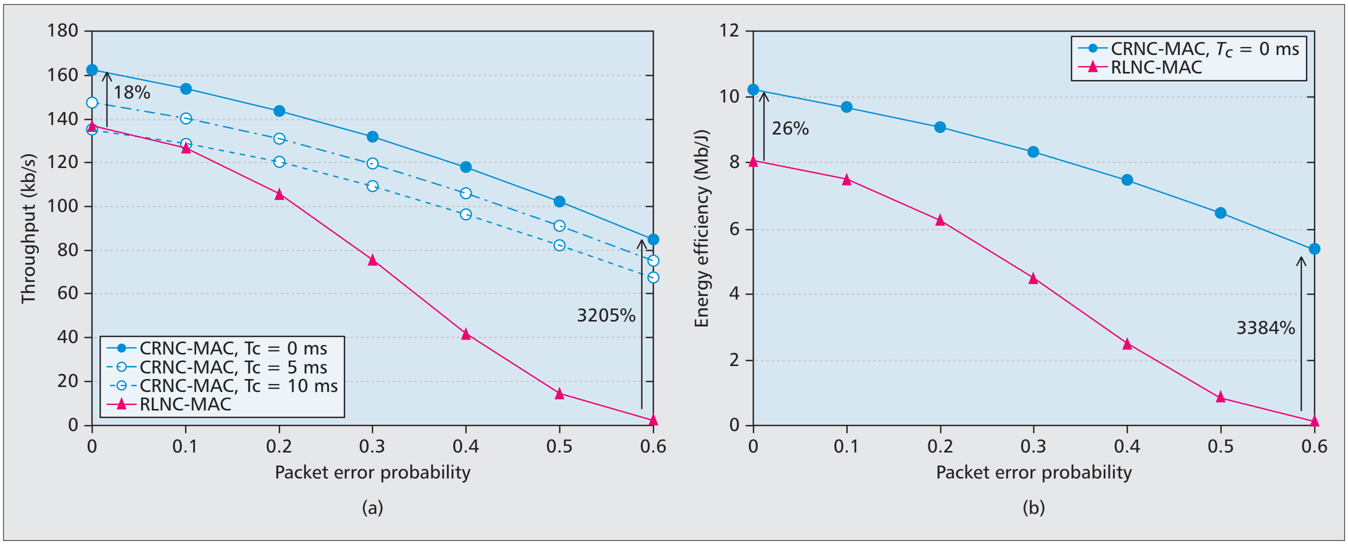


Figure 4. Performance evaluation with respect to the packet error probability p for $R = 3$ relays: a) throughput; b) energy efficiency.

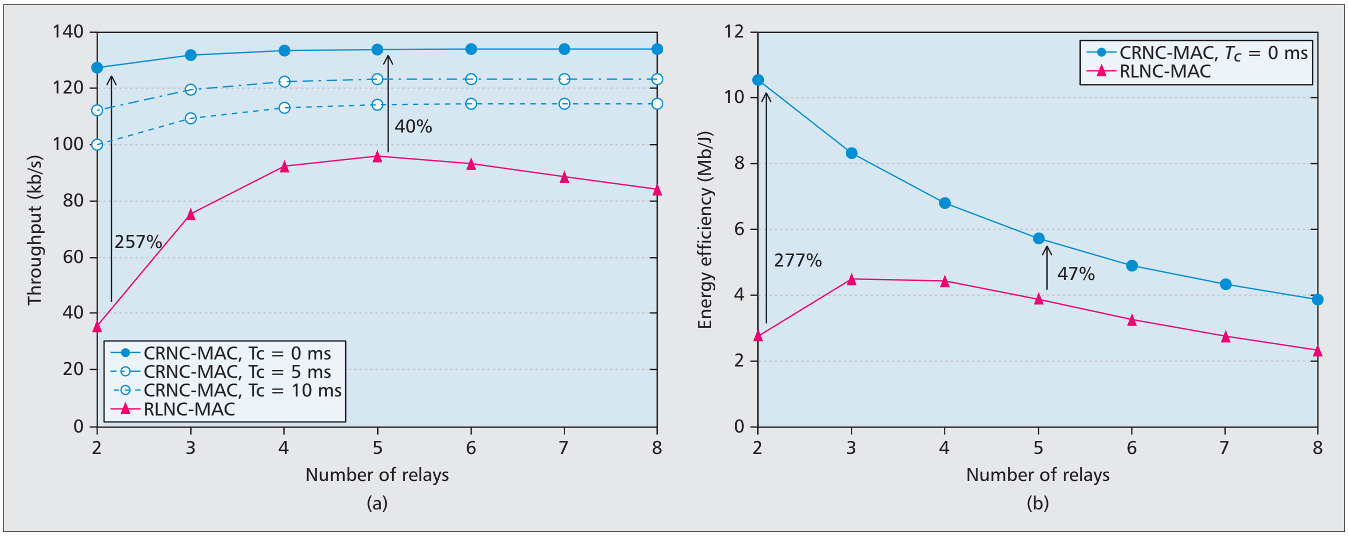


Figure 5. Performance evaluation with respect to the number of relays for a packet error probability of $p = 0.3$: a) throughput; b) energy efficiency.

plane within the relay cloud. The second approach proposes decentralized distribution of keys among users, but requires an arbitration entity for key recovery. In the proposed framework, the cloud controller can play the role of the arbitrator, facilitating key management operations such as key recovery and revocation.

Conclusions

In this article, we have presented a general framework for the design of efficient cloud-assisted protocols in AAL environments. The proposed framework defines two planes of operation for control and data, thus delegating network coordination tasks to a central entity located at the cloud, while all data related operations take place locally at the nodes.

Within this framework, we have developed a novel MAC protocol that manages to fully exploit the potential of RLNC in a cooperative relay network. The obtained results have shown significant performance gains that become more prominent under challenging scenarios when the channel conditions are harsh and only a few relays are available. Finally, we have also indicated other possible ways to exploit the potential gains offered by the cloud-assisted network coordination opening the road for many new applications.

Acknowledgments

This work has been funded by the research projects WSN4QoL (286047), SGR(2014 SGR 1551) and GEOCOM (TEC2011-27723-C02-01).

References

[1] UN, Dept. of Economic and Social Affairs, Population Division, "World Population Prospects: The 2012 Revision, Vol. I: Comprehensive Tables," ST/ESA/SER.A/336, 2013.

[2] P. Rashidi and A. Mihailidis, "A Survey on Ambient-Assisted Living Tools for Older Adults," *IEEE J. Biomed. Health Informatics*, vol. 17, May 2013, pp. 579–90.

[3] G. Van Den Broek, F. Cavallo, and C. Wehrmann, *AALIANCE Ambient Assisted Living Roadmap*, Amsterdam, the Netherlands: IOS Press, Jan. 2010.

[4] M. Alam, M. B. I. Reaz, and M. A. M. Ali, "A Review of Smart Homes — Past, Present, and Future," *IEEE Trans. Sys. Man Cybernetics C, Appl. Rev.*, vol. 42, Nov. 2012, pp. 1190–203.

[5] H. Pung *et al.*, "Context-Aware Middleware for Pervasive Elderly Homecare," *IEEE JSAC*, vol. 27, May 2009, pp. 510–24.

[6] A. Wood *et al.*, "Context-Aware Wireless Sensor Networks for Assisted Living and Residential Monitoring," *IEEE Network*, vol. 22, July 2008, pp. 26–33.

[7] M. Mulvenna *et al.*, "Visualization of Data for Ambient Assisted Living Services," *IEEE Commun. Mag.*, vol. 49, Jan. 2011, pp. 110–17.

[8] C. Bachmann *et al.*, "Low-Power Wireless Sensor Nodes For Ubiquitous Long-Term Biomedical Signal Monitoring," *IEEE Commun. Mag.*, vol. 50, Jan. 2012, pp. 20–27.

[9] R. Bassoli *et al.*, "Network Coding Theory: A Survey," *IEEE Commun. Surveys Tutorials*, vol. 15, 4th qtr. 2013, pp. 1950–78.

[10] E. AbuKhousa, N. Mohamed, and J. Al-Jaroodi, "e-Health Cloud: Opportunities and Challenges," *Future Internet*, vol. 4, no. 3, 2012, pp. 621–45.

[11] J. Wan *et al.*, "Cloud-Enabled Wireless Body Area Networks for Pervasive Healthcare," *IEEE Network*, vol. 27, Sept. 2013, pp. 56–61.

[12] J. H. Lim *et al.*, "A Closed-Loop Approach for Improving the Wellness of Low-Income Elders at Home Using Game Consoles," *IEEE Commun. Mag.*, vol. 50, Jan. 2012, pp. 44–51.

[13] A. Benharref and M. Serhani, "Novel Cloud and SOA-Based Framework for E-Health Monitoring Using Wireless Biosensors," *IEEE J. Biomed. Health Informatics*, vol. 18, Jan. 2014, pp. 46–55.

[14] P. Dely *et al.*, "CloudMAC — An OpenFlow Based Architecture for 802.11 MAC Layer Processing in the Cloud," *Proc. IEEE GLOBECOM Wksps.*, 2012, pp. 186–91.

[15] T. Uricar *et al.*, "Wireless-Aware Network Coding: Solving a Puzzle in Acyclic Multi-stage Cloud Networks," *Proc. 10th Int'l. Symp. Wireless Commun. Sys.*, Aug. 2013, pp. 1–5.

BIOGRAPHIES

ELLI KARTSAKLI received her Ph.D. in wireless telecommunications from the Technical University of Catalonia (UPC), Barcelona, Spain, in February 2012. She holds a degree in electrical and computer engineering from the National Technical University of Athens, Greece (2003), and an M.Sc. in mobile and satellite communications from the University of Surrey, United Kingdom (2004). She is currently a post-doctoral researcher at UPC, and has participated in several national and European projects (GREENET, WSN4QoL, etc.). Her primary research interests include wireless networking, channel access protocols, and energy efficient communication protocols.

ANGELOS ANTONOPOULOS received his Ph.D. degree (Cum Laude) from the Signal Theory and Communications (TSC) Department of the Technical University of Catalonia (UPC) in December 2012, and holds an M.Eng. degree from the Information and Communication Systems Engineering Department of the University of the Aegean (2007). He is currently working as a post-doctoral researcher, and his main research interests include cooperative communications, MAC protocols, network coding, and energy-efficient network planning. He has participated in several European and Spanish national projects (GREENET, Green-T, CO2GREEN, etc.) and has served as an expert evaluator of research projects funded by the Romanian Government through the National Council for Scientific Research. He has been granted three annual scholarships by the Greek State Scholarships Foundation (IKY) and recently has been awarded the First Polytechnic Graduates Prize by the Technical Chamber of Greece (TEE-TGC).

ARIS S. LALOS received a Ph.D. degree in signal processing for wireless communications from the Computer Engineering and Informatics Department (CEID), School of Engineering (SE), University of Patras(UoP), Rio-Patras, Greece, in 2010. He has been a research fellow at the Signal Processing and Communications Laboratory, CEID, SE, UoP from 2005 to 2010. In 2010–2011 he was a telecommunication research engineer at Analogies S.A, an early stage startup. He is currently a postdoctoral researcher in the Signal Theory and Communication D epartment, UPC.

STEFANO TENNINA (M.Sc 2003, Ph.D. 2007) was a research scientist in the CISTER Research Unit in Portugal from 2009 to 2012, involved in two projects: EMMON and SENODS. Since 2004 he has been with WESTAquila where he is currently a senior researcher involved in several EU and local projects. His research area is energy-efficient wireless communication protocols/systems. He has (co-)authored 30+ journal and conference papers. For more details, see http://www.tennina.net.

MARCO DI RENZO [SM'14] received a Ph.D. degree from the University of L'Aquila, Italy, in 2007. He is currently a "Charg de Recherche Titulaire" of the French National Center for Scientific Research (CNRS) at SUPELEC, France. He is a recipient of several best paper and personal awards, which include the 2013 IEEE ComSoc Best Young Researcher Award for the EMEA Region and the 2014 Royal Academy of Engineering Distinguished Visiting Fellowship. Currently, he serves as an Editor of *IEEE Communications Letters* and *IEEE Transactions on Communications*.

LUIS ALONSO obtained a Ph.D. from UPC in 2001 and got a permanent tenured position at the same university, becoming an associate professor in 2006. He was co-founder of the Wireless Communications and Technologies Research Group (WiComTec), to which he currently belongs. His current research interests are within the field of medium access protocols, radio resource management, cross-layer optimization, cooperative transmissions, cognitive radio, and QoS features for all kind of wireless communications systems. He is the author of 40 research papers in international journals and magazines, one book, 12 book chapters, and more than 100 papers at international congresses and symposia. He has received several best paper awards

CHRISTOS VERIKOUKIS got his Ph.D. from UPC in 2000. He is currently a senior researcher at CTTC and an adjunct professor at the University of Barcelona. His area of expertise is in the design of energy-efficient layer 2 protocols and RRM algorithms for short-range wireless cooperative and network coded communications. He has published 68 journal papers and over 140 conference papers. He is also a co-author of two books, 14 chapters in other books, and two patents. He has participated in more than 30 competitive projects, and has served as the principal investigator of national projects in Greece and Spain as well as the technical manager of Marie-Curie and Celtic projects. He has supervised 15 Ph.D. students and five postdoctoral researchers since 2004. He was General Chair of the 17th and 18th IEEE CAMAD, and TPC Co-Chair of the 15th Healthcom. He is currently serving as General Co-Chair of the 19th CAMAD and 6th IEEE Latincom. He is currently Secretary of the IEEE ComSoc Technical Committee on Communication Systems Integration and Modeling. He received the best paper award of the Communication QoS, Reliability & Modeling Symposium at IEEE ICC '11 and the EURASIP 2013 Best Paper Award for the *Journal on Advances in Signal Processing*.

GUEST EDITORIAL

EXTREMELY DENSE WIRELESS NETWORKS

Claudio Cicconetti

Antonio de la Oliva

David Chieng

Juan Carlos Zúñiga

The ever growing demand from mobile network users is pushing the current wireless technologies toward their limits. As a matter of fact, not even the most optimistic promises from emerging standards, such as LTE-Advanced and IEEE 802.11ac/ad/af, will be able to satiate the huge appetite for bandwidth of the future users of 5G networks — at least, unless the size of cells is reduced dramatically. This situation will soon create a desperate need for extremely dense wireless networks (DenseNets).

Unfortunately, merely shrinking the cell size is not sufficient due to several technical and economic factors. On the technical side we find interference, which is a real challenge as the number of cells deployed nearby increases, and energy consumption, which also becomes important due to much more irregular usage patterns. Among the business/deployment issues, we mention the availability of sufficient backhaul capacity everywhere, the current lack of very low-cost/very small base stations, and the need for economically viable operation, administration, and management (OAM) activities.

In brief, low-hanging fruits will not be there to be grabbed by the telecom industry. Rather, a new ecosystem of solutions, possibly characterized by the use of cooperative approaches, will be required to take full advantage of the new opportunities brought by extremely dense wireless networks. The nine articles in this Feature Topic deeply investigate some of the hottest research challenges in this context.

The first challenge addressed is related to the deployment of future mobile networks. In "Spectral and Energy Efficiency of Ultra-Dense Networks under Different Deployment Strategies" by Syed Fahad Yunas *et al.*, it is shown how extremely dense indoor femtocells and outdoor distributed antenna systems both outperform a densified macrocell deployment in terms of capacity and energy efficiency. Therefore, to meet the 5G requirements, alternative paths to the long-standing processes of cellular network provisioning must be experimented.

The second article, "Interference Coordination for Dense Wireless Networks" by Beatriz Soret *et al.*, presents the problem of interference in dense scenarios: as the density of cells increases, so does interference. This article revisits the options on the table for LTE and LTE-Advanced, proposing two algorithms to apply time domain and frequency domain small cell interference coordination in a DenseNet.

The third article, "Understanding Channel Selection Dynamics in Dense Wi-Fi Networks" by Akash Baid and Dipankar Raychaudhuri investigates the impact of increasing enterprises' or service providers' access points (APs) with centralized channel assignment on the performance of typical residential APs and vice versa. A parametric approximation scheme is proposed for estimating the AP's throughput.

The fourth article, "Per-Node Throughput Enhancement in Wi-Fi DenseNets" by Kyungseop Shin *et al.*, proposes a joint dynamic sensitivity control (DSC) and transmit power control (TPC) approach to control interference in dense residential, enterprise, or indoor hotspot deployments. The simulation results show that significant per-node throughput improvements can be achieved regardless of the deployment type.

The fifth article, "On the Efficient Utilization of Radio Resources in Extremely Dense Wireless Networks" by Arash Asadi *et al.*, tackles the joint utilization of WiFi relays to improve the performance of LTE. This work proposes a resource allocation mechanism opportunistically exploiting network density as a resource. Results show that intracell opportunistic relay can reduce the complexity and boost efficiency of intercell interference coordination in LTE.

The sixth article, "Toward 5G DenseNets: Architectural Advances for Effective Machine-Type Communications over Femtocells" by Massimo Condoluci *et al.*, presents a novel architecture to handle the growing MTC traffic by the use of small cells to handle the massive and dense MTC rollout. This work introduces a novel 3GPP-compliant architecture that absorbs the MTC traffic via home evolved NodeBs (HeNBs), providing significant reduction of congestion in radio access and core networks.

Efficient mobility management is a firm but challenging requirement in future extremely dense scenarios of mobile networks. In the seventh article, "Distributed Mobility Management for Future 5G Networks: Overview and Analysis of Existing Approaches" by Fabio Giust *et al.*, the authors propose a novel and highly scalable mobility management architecture based on the distributed mobility management concept currently under study at the Internet Engineering Task Force (IETF).

C-RAN is a promising technology to increase the density of the network at reduced cost. This topic is investigated in the eighth article, "Software-Defined Networking in Cellular Radio Access Networks: Potential and Challenges" by Mustafa Y. Arslan *et al.*, where the concept of SDN is applied to the fronthauling in C-RAN deployments. The results can also be applied to extended support for coordinated multipoint approaches.

Finally, the ninth article, "Scalability of Dense Wireless Lighting Control Networks" by Conrad Dandelski, addresses an interesting technological alternative for DenseNets, that is, the use of LED-based lighting networks already deployed in buildings for broadcasting messages and collecting sensor data, thus providing a control system for the lighting network increasing its scalability.

ACKNOWLEDGMENT

Putting together this Feature Topic has been challenging but rewarding. We would like to thank all authors for their invaluable contributions. This Feature Topic would not be possible without the timely and professional support of all the reviewers. Last but not least, we would also like to record our sincere thanks to Dr. Sean Moore, Editor-in-Chief of *IEEE Communications Magazine*, Charis Scoggins, and Jennifer Porcello for their relentless support throughout the whole publication process.

BIOGRAPHIES

CLAUDIO CICCONETTI (ccicconetti@mbigroup.it) holds a Laurea degree (2003) and a Ph.D. (2007) in information engineering from the University of Pisa. He worked at Intecs as an R&D Manager until 2013 and is now working at MBI S.rl. (Italy) on satellite communications. He has been involved in several projects co-funded by the European Commission (EuQoS, SANDRA, CROWD, MOTO, BETaaS) and the European Space Agency (SAT4NET, ACCORD). He has co-authored 50+ scientific papers and two international patents.

ANTONIO DE LA OLIVA (aoliva@it.uc3m.es) obtained a degree (2004) and his Ph.D. (2008) in telecommunication engineering at the University Carlos III of Madrid, where he is an assistant professor. He has been involved in many European projects (DAIDALOS, ONELAB, CARMEN, MEDIEVAL, CROWD). He has published 30+ papers and submitted several patents on mobility management and wireless networks. He has served as Vice-Chair of the IEEE 802.21b task group and Technical Editor of IEEE 802.21d.

DAVID CHIENG (ht.chieng@mimos.my) is currently the head of the HetNet Lab in Mimos, Malaysia. From 2006 to 2010, he led broadband wireless research in BT MRC and also served as a WP manager in the EU FP7 CARMEN project. From 2003 to 2006, he was with the Faculty of Engineering, Multimedia University. He received his Ph.D. in communication networks from Queen's University of Belfast, Northern Ireland, in 2002. He has authored and co-authored more than 40 publications and filed 14 patents.

JUAN CARLOS ZÚÑIGA (JuanCarlos.Zuniga@interdigital.com) received his engineering degree from the UNAM, Mexico, and his M.Sc. from Imperial College London, United Kingdom. He has worked with Harris Canada, Nortel Networks UK, and Kb/Tel Mexico, and is now a principal engineer at InterDigital. He chairs the IEEE 802 Privacy SG and IETF Internet Area WG. He is the inventor of 30+ granted patents and has several publications. He has been granted the BoD Chairman's Award and the British Council Fellowship.

Spectral and Energy Efficiency of Ultra-Dense Networks under Different Deployment Strategies

Syed Fahad Yunas, Mikko Valkama, and Jarno Niemelä

Syed Fahad Yunas and Mikko Valkama are with Tampere University of Technology.

Jarno Niemelä is with Elisa Corporation.

ABSTRACT

To tackle the 1000× mobile data challenge, the research towards the 5th generation of mobile cellular networks is currently ongoing. One clear enabler toward substantially improved network area capacities is the increasing level of network densification at different layers of the overall heterogeneous radio access system. Ultra-dense deployments, or *DenseNets*, seek to take network densification to a whole new level, where extreme spatial reuse is deployed. This article looks into DenseNets from the perspectives of different deployment strategies, covering the densification of the classical macro layer, extremely dense indoor femto layer, as well as outdoor distributed antenna system (DAS), which can be dynamically configured as a single microcell or multiple independent microcells. Also, the potential of a new indoor-to-outdoor service provisioning paradigm is examined. The different deployment solutions are analyzed from the network area spectral and network energy efficiency perspectives, with extreme densification levels, including both indoor and outdoor use scenarios. The obtained results indicate that dedicated indoor solutions with densely deployed femtocells are much more spectrum- and energy-efficient approaches to address the enormous indoor capacity demands compared to densifying the outdoor macro layer, when the systems are pushed to their capacity limits. Furthermore, the dynamic outdoor DAS concept offers an efficient and capacity-adaptive solution to provide outdoor capacity, on demand, in urban areas.

INTRODUCTION

The global deployment of 4G/IMT-Advanced networks is still in its early phase, with major deployments expected to commence in 2015. However, with the exponentially increasing global data traffic volume together with a projected massive increase in the number of connected devices in the near future, it is envisioned that the capacity of fourth generation (4G) networks may already reach their limits soon. According to recent predictions, the amount of global data traffic increased more than twofold between 2010 and 2011 [1]. With this annual growth rate, industry experts have predicted a significant 1000× increase in the total data capacity demand in the next 10 years. As a preemptive solution, to deflect the danger of running into a capacity crunch, the mobile industry is already working toward the fifth generation (5G) of wireless cellular networks, which is conceived to address the growing capacity demand in a sustainable and cost-effective manner with substantially lowered energy consumption per transferred bit. 5G networks will not be just about enhancements of the radio access network (RAN), but rather will represent an ecosystem of interoperable technologies and network layers, working as a whole to provide ubiquitous high-speed connectivity.

To tackle the 1000× Data Challenge, as some industry leaders call it [1], network vendors and mobile operators have to focus on two partially related key aspects:

- High-bit-rate service provisioning
- Ubiquitous — *anywhere anytime* — service provisioning

The first strategy is the traditional approach, the industry has followed up to now (i.e., increasing the cell level capacities by improving the air interface efficiency through *advanced digital transmission techniques* (higher order modulation and coding, advanced antenna systems etc.,) and *utilization of larger spectrum chunks*. Although such improvements at the air interface significantly improve the cell level capacities, they are still not able to provide the needed network level gains. Hence, a very different approach is needed at the system level to meet the imminent explosive growth in data traffic demands.

The second strategy focuses on providing ubiquitous anywhere any time service to the masses (i.e., increasing the network level capacity to support more users and devices in a given area). One of the most obvious ways to increase the capacity of a wireless network is by spatially reusing the existing allocated spectrum as frequently as possible throughout the network service area, in other words, increasing the base station density. As such, the capacity of a cellular network is considered to be proportional to

0163-6804/15/$25.00 © 2015 IEEE

the base station density. This article discusses *DenseNets* from an alternative deployment strategies perspective, in particular when individual densification solutions are pushed to their capacity limits. We start by looking into a coventional methodology of network densification used by operators, mostly based on macrocellular densification, discussing the limitations of that approach, and then proceed toward newer deployment paradigms that enable successful realization of *DenseNets* concepts. We particularly emphasize in our analysis and presentation *network-level spectrum efficiency* as well as *network energy efficiency* when different deployment solutions, in this case densified macro networks, extremely dense small cell networks, and distributed antenna system (DAS)-based networks, are pushed to the limits such that all the network elements operate at full load. We also pay special attention to the differences between indoor and outdoor user equipments (UEs) under these different deployment solutions, strongly motivated by recent observations (e.g., in [2]) that wall penetration losses of both residential and commercial buildings can peak up to 30 dB or so due to new construction materials with high thermal insulation, also impacting radio signal propagation. Before going into the deployment-level presentation, the analysis methodology for the performance evaluation of different deployment strategies, discussed in the subsequent sections of this article, is herewith briefly covered. Further details on the analysis methodology and simulation parameters can be found in [4–6].

Analysis Methodology

The key assumptions and tools we use in the performance analysis of different deployment and densification solutions are highlighted as follows:

- For modeling the outdoor and indoor radio channels, a deterministic ray-tracing-based radio propagation model is deployed.
- A homogeneous propagation environment is assumed (i.e., all the cell sites experience similar radio propagation conditions). As such, the dominance areas of all the cell sites are identical. Hence, for the performance evaluation of different deployment strategies, the receiver points from the dominance (*best server*) area of the center cell site are considered for statistical analysis, while other cells or transmission points are treated as interference.
- In the case of DAS configuration, the received signal power from a serving DAS cell is actually the superposition of the received signal powers from all the individual nodes belonging to the serving DAS cell; that is, the signal powers from all the nodes within a DAS cell are combined at the receiver, while the signal power received from other nodes is treated as interference.
- For simulating a continuous cellular network effect, the dominant interfering tiers contributing significantly to the interference level in the dominance area of a serving cell are considered.
- The distribution of receiver points outdoors and across all the buildings (floors) is uniform.
- Full cell load over the network is assumed, which is the worst case scenario)i.e., all the cells are transmitting at full power), by which we seek to push different deployment solutions to their ultimate limits in a systematic manner.
- In order to evaluate the full potential of the different deployment strategies, no backhaul limitation is assumed. Furthermore, classical Shannon information theoretic laws of the form $\log_2(1 + SINR)$ are used to map experienced signal-to-interference-plus-noise ratios (SINRs) to cell-level spectral efficiencies.

The metrics used for the performance analysis are briefly described below:

Cell and Network Spectrum Efficiency — The *cell spectrum efficiency* (bits per second per Hertz) is estimated using the Shannon capacity bound and averaged within cell across all receiver points. The network spectrum efficiency (NSE) is then the actual network -level spectrum efficiency measure, which depends on the spatial reuse over a given area. As a result of cell network densification, the spatial reuse within an area increases, which results in network-level capacity gain. In other words, the network capacity depends on the cell density within a given area. In our analysis, we normalize the network-level capacity over 1 Hz of bandwith and over an area of 1 km^2. Mathematically, *NSE* [*kb/s/Hz/km*^{2}] = (*cell spectrum efficiency*) × (*cell density over 1 km*^{2}).

As a result, for cell network densification, the spatial reuse within an area increases, which results in network-level capacity gain. In other words, the network capacity depends on the cell density within a given area.

Energy Efficiency — The energy efficiency, E_{eff}, of a network is defined as the aggregate bit rate that is achievable over 1 Hz nominal bandwidth while consuming a given power, e.g., 1 kW — thus measured in bits per second per Hertz per kilowatt. This methodology is appropriate for assessing the energy efficiency of a network operating under full load condition [3]. As such, the energy efficiency of a network is given by

$$E_{eff} = \frac{NSE}{network\ power\ consumption\ over\ 1\ km^2}.$$

Table 1 gathers the essential system parameters used in the performance analysis of different deployment strategies. We emphasize that these selected parameters represent realistic and typical example scenarios from the deployment, propagation, and device perspectives.

Related Work in the Literature

Numerous studies with wide ranging scope related to network densification have been reported in the past; for example, the impact of base station densification on the capacity and energy efficiency on different base station types were reported in [7, 8]. The total power consumption of different network densification alternatives in the Long Term Evolution (LTE) context was studied in [9], which concluded that under low discontinuous transmission (DTX), macrocell densification is the most power-efficient solution. The impact of inter-cell interference, resulting from densifying outdoor small cells, on the user experience is investigated in [10]. The authors aim to investigate *if and under what conditions* intercell coordination is useful as com-

Contemplating the relative share of today's indoor/outdoor data traffic, there is a global consensus among mobile operators and wireless infrastructure vendors that majority of the data traffic, approximately 65–70 percent, is generated by indoor users.

Parameter	Unit	Value
Operating frequency	MHz	2100
Bandwidth, W	MHz	20
Macrocell densities	Cells/km^2	3.8, 5.1, 9.9, 39.3, 119.9
Indoor femtocell densities	Cells/km^2	4125 (suburban), 5197 (urban)
Outdoor DAS nodes densities	Nodes/km^2	145
Transmit power	dBm	Macrocell: 43, DAS nodes: 30, femtocell: 20
BS antenna type		Directional (macro), Omni (DAS nodes and FAP)
BS antenna beamwidth, $HPBW_{h/v}$	°	Directional (65°/6°), omni (360°/90°)
BS antenna gain, G_m	dBi	Macrocell: 18, DAS node: 5, femtocell: 2.2
UE antenna type		Halfwave dipole
UE antenna gain	dBi	2.2
BS antenna height, h_{BS}	m	Macrocell: 30, DAS node: 8, femtocell: 2 m above indoor floor levels
UE antenna height (above floor level), h_{MS}	m	2
Receiver noise figure	dB	9
Receiver noise floor level, P_n	dBm	–92
Propagation model		Dominant path loss model

Table 1. General evaluation parameters.

pared to universal frequency reuse. The conclusions drawn from the paper are that in street deployed small cells, the intercell interference does not impact user throughput; hence, no inter-cell coordination is required; intercell coordination is helpful only in open areas, where intercell interference between small cells is significant. In [11], a novel architecture based on software defined networking is proposed to address the various challenges arising from deploying DensenNets (mobility management, network optimization, etc.). Another promising technology for improving the coverage and capacity of future wireless systems is the distributed antenna system (DAS). The information theoretical ergodic capacity of DAS has been studied in some recent works, such as [12, references therein]. In [12], the authors evaluate and analyze the ergodic capacity of downlink DAS with random node placement using a stochastic-geometry-based method. The numerical results from the study indicate that the DAS has better cell edge performance than traditional wireless network.

The next sections in this article look into the performance of different deployment methodologies, starting with the classical deployment strategy based on macrocellular densification, including both outdoor and indoor users. We then look into the performance of *ultra dense indoor small cells* in suburban and urban environments. Particular emphasis is given to the impact of varying wall penetration losses on the performance of ultra dense small cells, from the indoor and potential indoor-to-outdoor service provisioning perspectives. We introduce the *dynamic DAS* concept as an effective solution for providing on-demand cell and system-level capacity based on variable outdoor data traffic. Then concluding remarks are given.

Classical Deployment Strategy: Outdoor Macrocellular Network

In the past and still today, mobile operators have built their networks using the *outside-in* approach (i.e., relying primarily on outside macro base stations). Looking at the network evolution, initially, the network is designed from the coverage perspective by deploying macrocell sites to serve both outdoor and indoor locations with a certain minimum quality of service. As the number of devices accessing the network increases, it transitions from coverage limited to capacity limited state, thus necessitating denser deployments. The densification of the network is done gradually. In the initial stages, the mobile operator tries to accommodate the network capacity demands by densifying the macro layer itself by installing more macro base stations. As the network matures, and the number of devices accessing the services keeps increasing, several capacity-limited local hotspot areas within the network begin to appear. These hotspots, limited in size and scattered throughout the network service area, can then be covered by the operators by deploying street-level microcells, thus forming what is typically known as a *hierarchical cellular structure* in which macrocells provide the umbrella coverage and microcells aim to fulfill the capacity demands in local hotspot areas.

However, as the demand for further capacity increases (mostly coming from indoor locations) the achievable network capacity from densifying the outdoor layers begins to saturate, and operators are forced to transition toward indoors (i.e., start deploying dedicated indoor solutions). Contemplating the relative share of today's indoor/outdoor data traffic, there is a global consensus among mobile operators and wireless infrastructure vendors that the majority of data traffic, approximately 65–70 percent, is generated by indoor users. Assuming that this ratio of outdoor/indoor traffic share will persist in future as well, the projected capacity demand from indoor users can then be estimated to increase approximately 650–700× (assuming 1000× overall increase in data traffic demand). Such a massive increase in capacity demand from indoor users cannot be delivered in a cost- and energy-efficient manner by dense outdoor deployments alone due to associated indoor capacity inefficiencies. Furthermore, poor indoor coverage from outdoor deployments in buildings with high penetration loss has been and still is the topmost complaint with which mobile operators are struggling. Apart from the capacity and coverage limitations, the outdoor network densification also suffers from energy inefficiency, a key concern for mobile operators.[1]

As an example, by deploying the analysis methodology described earlier, Fig. 1 shows the performance of a macrocellular network densification from the cell spectrum efficiency, network spectrum efficiency, and network energy efficiency points of view for both outdoor and indoor user locations in an urban environment using the applicable evaluation parameters in Table 1. A hexagonal layout has been used as the basis for macrocellular deployment, and the network is deployed in a Manhattan type grid, as shown in Fig. 1. It is clear that the indoor environment suffers from capacity inefficiency with increasing cell density. As the network is densified, the sites are brought closer together, which increases the intercell interference. In order to limit the intercell interference in the macro layer, the antennas have to be down-tilted to a greater degree [4]. This results in poor coverage on some floor levels inside high rise buildings, which degrades the radio channel conditions, thereby affecting the overall achievable cell spectrum efficiency. Furthermore, as a result of reduction in cell-level spectrum efficiency, the network spectrum efficiency starts to saturate in the indoor environment and hence implies that macrocellular network densification becomes less efficient from the indoor service provisioning point of view. From the network energy efficiency perspective, the indoor capacity inefficiency has a direct impact on the energy efficiency performance of the network, as lower spectrum efficiency in the indoor environment results in higher energy consumption per bit, in other words, lower network throughput per unit power consumed.

The fundamental behavior seen in Fig. 1 necessitates a shift from the current outside-in approach to a new deployment paradigm that puts more focus on service provisioning from the indoor perspective. As such, indoor deployment of small cells has been identified as a cost-efficient solution that offers wireless carriers a sustainable evolutionary pathway to meet indoor capacity demands in the future [13, 14]. An increasing trend among operators opting for such small cell deployments has recently been observed.

For Beyond 4G (B4G) networks, the experts envision that to fulfill the surging capacity demands of 1000× or more, an extremely dense network of small cells which provides seamless coverage and mobility, thus giving rise to a concept known as DenseNets. Network densification, based on ultra-dense deployment of small cells, is being considered as one of the key flavors of the emerging 5G cellular networks that will truly address the 1000× data challenge [13, 14]. A large share of these deployments will be indoors, as this is the arena from which the majority of data traffic is expected to originate in the future. While the basic outdoor macro layer is still always needed for high mobility outdoor users, such massive scale indoor deployments may also shift the current outside-in deployment strategy toward a new paradigm based on an inside-out approach where not only indoor users but potentially some low-mobility outdoor neighborhood users as well can be served by indoor base stations [15]. These are explored next.

While the outdoor basic macro layer is still always needed for high mobility outdoor users, such massive scale indoor deployments may also shift the current Outside-In deployment strategy towards a new paradigm based on Inside-Out approach.

Indoor and Indoor-to-Outdoor Service Provisioning

As mentioned in the previous section, 5G networks will most likely encompass extremely dense deployments of small cells, typically in indoor scenarios. It is believed that the of these small cells will be purchased and deployed by either end users in their homes in a "plug and play" fashion, or enterprises in commercial buildings with no or minimal assistance from mobile operators, thereby enabling significant savings for the operators in terms of capital expenditures (CAPEX) and OPEX. Moreover, due to a very small coverage footprint, the extreme density of these small cells will enable very tight frequency reuse, resulting in large *network capacity* gains, thus fulfilling the indoor capacity demands in a cost-efficient manner.

For outdoor service provisioning, the operators have so far been relying on outdoor installations (mostly macro layer), while deploying microcells only in hotspot areas. Outdoor deployments, unfortunately, come with a sub-

[1] Recently, political initiatives have put stringent requirements on mobile operators to reduce their CO_2 emissions. Also, from the cost perspective, the energy consumption of radio access networks contribute significantly toward mobile operator operational expenditures (OPEX). Roughly 80 percent of the energy consumed by the RAN comes from the base stations.

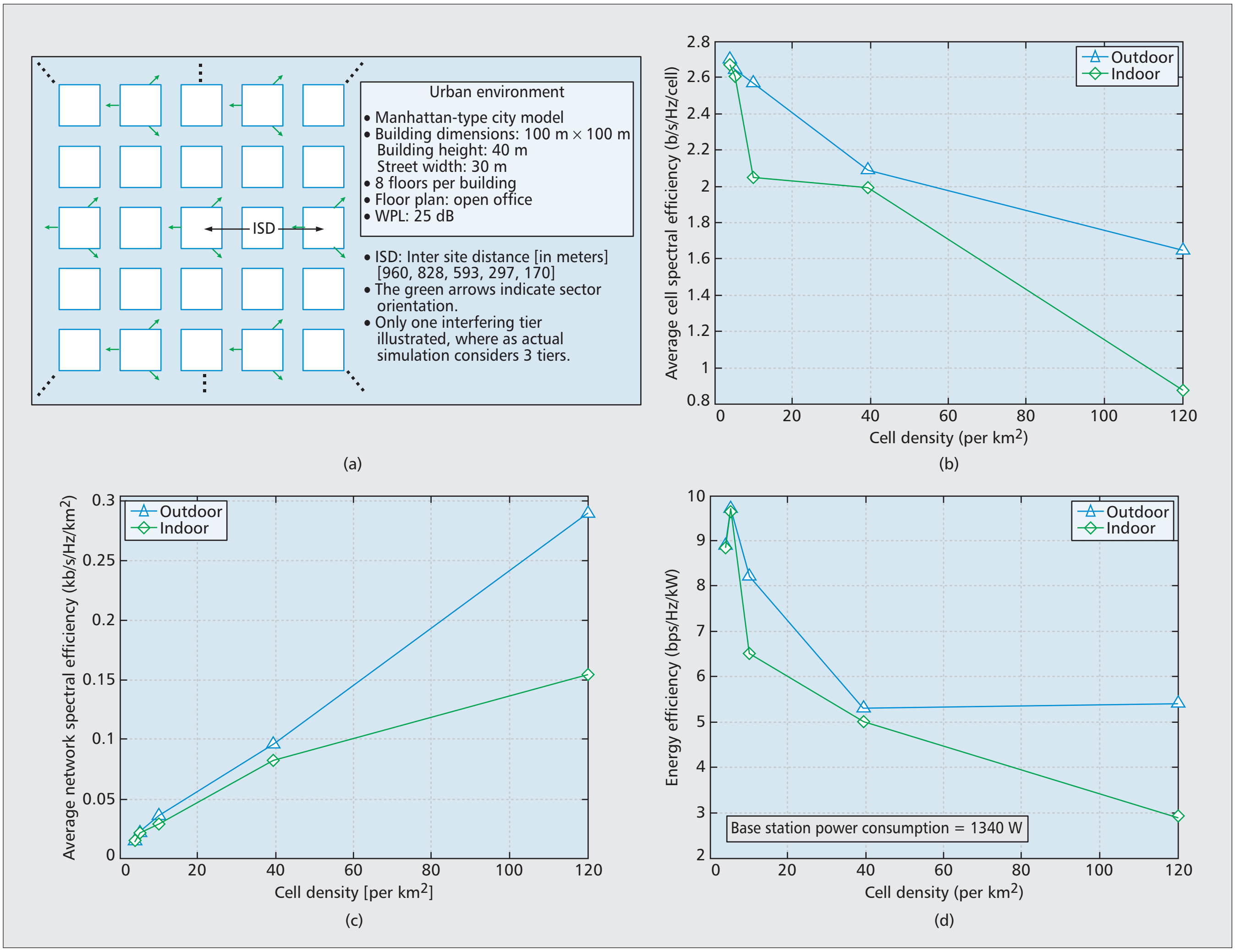


Figure 1. Performance analysis of macrocellular densification from the outdoor and indoor receiver perspectives: a) Manhattan grid model (aerial view); the green arrows show sector antenna positions and orientations (example case: ISD 297 m); b) average cell spectrum efficiency; c) average network spectrum efficiency; d) network energy efficiency vs. cell density.

stantial price tag for wireless carriers, with major cost items being site rental, RF engineering, and backhaul connectivity. The associated high CAPEX and OPEX, nowadays, are a key concern for cost-aware mobile operators striving to provide better services at lower cost in a highly competitive market, where the flat business model prevails. Fortunately, there is an option for mobile operators to bring their infrastructure cost down significantly. Indoor small cells, despite having limited coverage footprint, tend to radiate/spill their signals into the neighborhood outdoor environment. These signals usually originate from small cells located in nearby buildings. By enabling the indoor small cells to operate in an open subscriber group (OSG) mode and thus provide service in their immediate outdoor vicinities, mobile operators can significantly lower their infrastructure costs by benefiting from zero *site rental*, *RF engineering*, and *backhaul connectivity* costs, thereby providing connectivity to outdoor users/customers with lower incurred costs. A similar concept of indoor-to-outdoor service provisioning has been presented by Qualcomm as neighborhood small cells (NSC) in [13, 15]. Furthermore, key challenges related to deployment, mobility management, and radio resource management (RRM) of NSC were discussed quite well in [15]. However, one key item that is still missing from the studies is how the indoor-to-outdoor service provisioning (IOSP) concept will perform in modern construction with high wall penetration losses (WPL), reported, for example, in [2].

Recently, due to the increased level of awareness of global warming, and the resulting requirements to save energy and cut down CO^2 emissions, the construction industry has started to develop, manufacture, and utilize modern construction materials that provide a greater degree of thermal insulation. Unfortunately, these types of materials significantly impact radio propagation in the form of high building penetration loss (BPL). Traditionally, the values have been in the range of 5–15 dB; however, a more recent study has reported building penetration losses up to 35 dB in modern constructions [2]. High wall penetration losses attenuate signals penetrating through them, resulting in signal quality deterioration, which in turn affects

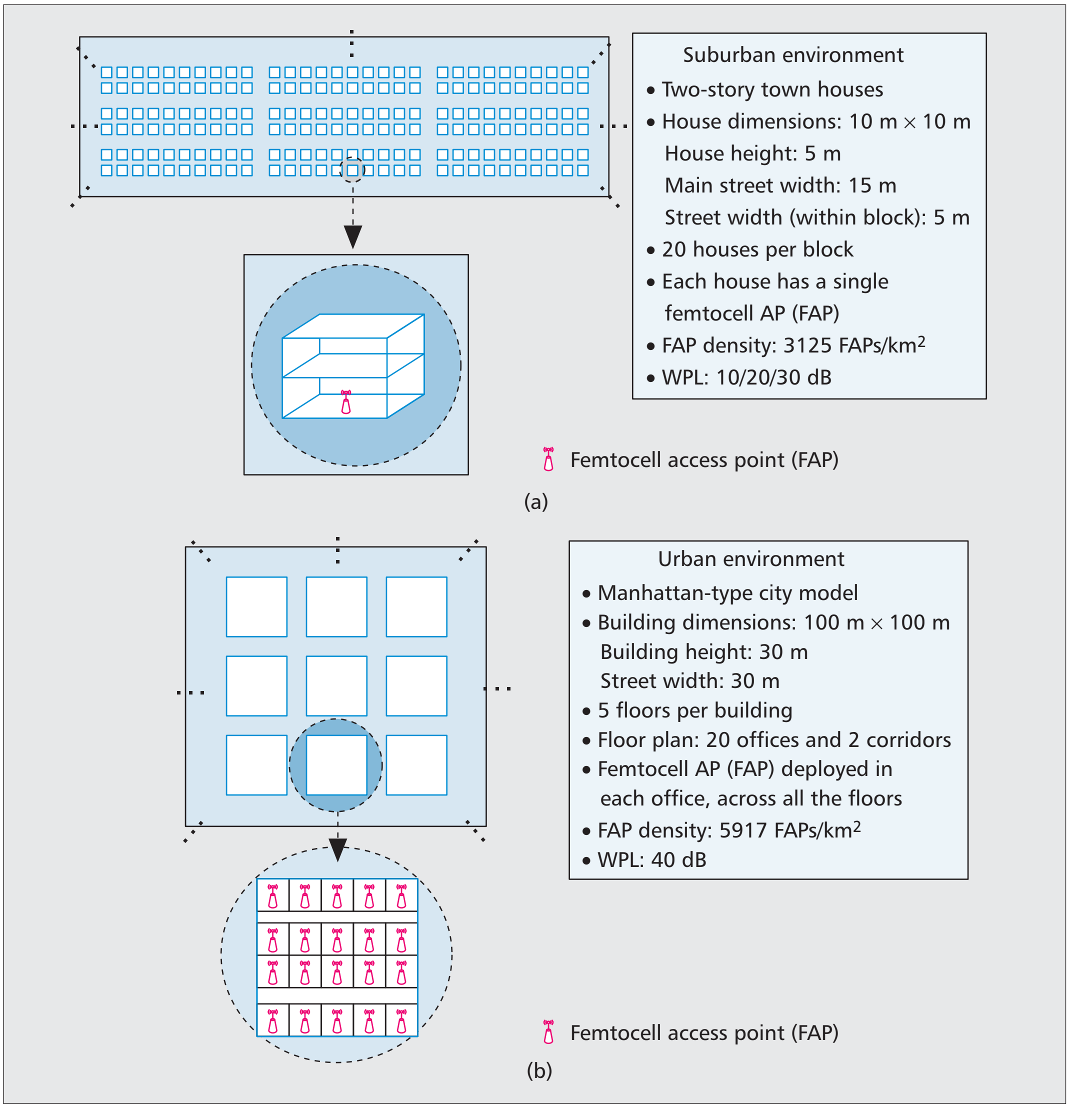

Suburban environment
• Two-story town houses
• House dimensions: 10 m × 10 m
House height: 5 m
Main street width: 15 m
Street width (within block): 5 m
• 20 houses per block
• Each house has a single femtocell AP (FAP)
• FAP density: 3125 FAPs/km²
• WPL: 10/20/30 dB
Femtocell access point (FAP)
(a)
Urban environment
• Manhattan-type city model
• Building dimensions: 100 m × 100 m
Building height: 30 m
Street width: 30 m
• 5 floors per building
• Floor plan: 20 offices and 2 corridors
• Femtocell AP (FAP) deployed in each office, across all the floors
• FAP density: 5917 FAPs/km²
• WPL: 40 dB
Femtocell access point (FAP)
(b)

Figure 2. Indoor femtocell-based DenseNets scenarios: a) suburban environment (aerial and 3D view); b) urban Manhattan type environment (aerial view only).

High wall penetration losses attenuate signals penetrating through them, resulting in signal quality deterioration, which in turn affects the network capacity and data throughput. Hence, we consider the impact of different wall penetration losses on the performance evaluation of indoor DenseNets.

network capacity and data throughput. Hence, we consider the impact of different wall penetration losses on the performance evaluation of indoor DenseNets from the indoor and indoor-to-outdoor service provisioning perspectivea in our analysis. This is addressed next.

Figure 2 shows two different environments, suburban and urban, where the performance of indoor DenseNets with extremely densely populated femto access points (FAPs) in an OSG configuration is evaluated. In OSG mode, outdoor macrocell users can also attach to indoor femtocells in locations where the indoor femtocells are dominant and vice versa. In the suburban scenario, we consider three different wall penetration losses: 10, 20, and 30 dB. The values have been selected to model the exterior wall penetration losses of typical town houses with older and modern constructions [2]. In turn, for the urban environment, we assume a downtown (city center) area with tall buildings. The external wall penetration loss in such urban buildings is believed to be even higher due to the presence of steel columns and so on; hence, the WPL in this case is assumed to be 40 dB. Furthermore, as we are evaluating the full potential of indoor DenseNets in both suburban and urban environments, a full network load is assumed, that is, all the cells are transmitting at full power (limiting case scenario) with no backhaul limitation. From the deployment point of view, in a suburban environment, a single femtocell is assumed per residential home, which results in a cell density of 3125 FAPs/km^2. In the urban environment, due to larger buildings, 20 FAPs per floor are assumed (1 FAP in every open office), which gives an extremely high cell density of 5917 FAPs/km^2.

Figure 3 shows the spectrum efficiency and energy efficiency performance for indoor deployed DenseNets in suburban and urban environments for different WPL scenarios. From the indoor users' perspective, the high WPL is actually shown to improve the capacity performance in the suburban environment. The reason is simply due to better shielding, with increasing WPL, from neighboring interfering FAPs that are installed in other houses, as shown also by

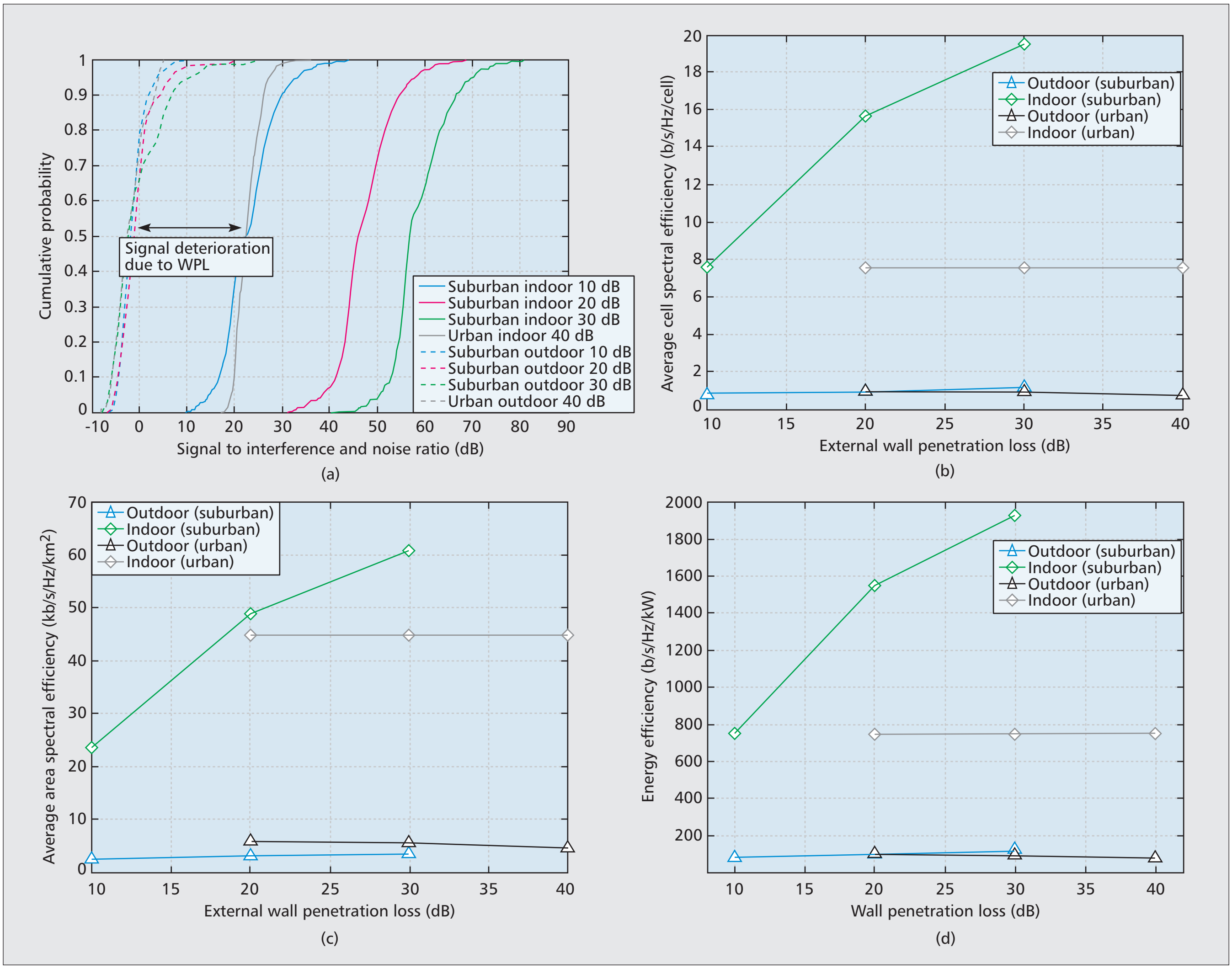


Figure 3. Performance analysis of femtocell-based DenseNets for outdoor and indoor users in suburban and urban environments with different wall penetration losses: a) SINR (dB); b) average cell spectrum efficiency (b/s/Hz/cell); c) average network spectrum efficiency (kb/s/Hz/km^2); d) network energy efficiency (b/s/Hz/kW).

the SINR performance in Fig. 3a. Due to very high frequency reuse (resulting from ultra dense indoor deployments), along with better interference shielding due to high WPL yielding high cell-level capacities, the area capacity in the indoor environment increases proportionally. For the urban environment, on the other hand, due to the dominant interferers being present inside the building, the cell level capacities are deteriorated, resulting in lower cell spectrum efficiency compared to suburban scenarios. Furthermore, the lower cell level capacity in the urban scenario also naturally decreases the achievable area capacity. Hence, we see slightly lower network spectrum efficiency performance in the urban environment than in the suburban 20 and 30 dB WPL scenarios. For comparison purposes, WPLs of 20 and 30 dB are also plotted for the urban scenario. In the indoor environment, as the dominant interferers are already present inside the building, the wall penetration loss does not have any essential impact on the cell, network, and energy efficiency performance.

Looking next at the indoor-to-outdoor service provisioning performance, the achievable outdoor capacity gain from dense indoor femtocell deployments is comparatively lower. The reason, as mentioned earlier, is attributed to the high WPL, which actually deteriorates the signal quality of indoor transmitters, as clearly shown in Fig. 3a, thereby reducing cell throughput. Comparing the network spectrum efficiency and network energy efficiency with densest macro-layer configuration (120 cells/km^2), the indoor ultra dense femtocell deployment offers 400x more indoor area capacity and 12× more outdoor area capacity in suburban 30 dB WPL scenario, while in the urban environment, the relative indoor/outdoor area capacity gain is 150× and 16×, respectively, in the 40 dB external wall penetration scenario. Similarly, from the energy efficiency point of view, due to the extremely high frequency reuse and significantly low power consumption per FAP (on the order of 10 W), the amount of bits transferable per kilowatt in the suburban 30 dB scenario is 600× and 20× more than in the densest macrocellular configuration for indoor and outdoor, respectively.

Hence, we can conclude that for indoor users, dense femtocell deployment provides enormous network level capacity gain, and also reduces the energy per communicated bit substantially compared to a densified macro network. Furthermore, mobile operators can provide certain services to outdoor users from indoor access points; however, in order to guarantee higher bit rates, the operators will need to deploy dedicated outdoor installations as well. Furthermore, the indoor-to-outdoor service provisioning will obviously work only in small streets or neighbourhoods, as the indoor small cells, due to lower transmit power levels, will only be able to cover areas in the vicinity of the buildings. Any outdoor location wider than a few tens of meters might experience coverage limitation if there is no dedicated outdoor access layer available. An indoor-to-outdoor service provisioning solution using indoor DenseNets can thus be considered as a good complement to the outdoor network, as a means of, say, offloading capacity in times when the outdoor layer is overloaded with users during busy hours.

Outdoor Dynamic Distributed Antenna Systems

So far we have focused on the service provisioning from the indoor capacity demand perspective, and also looked at deployment concepts that would be useful not only for providing substantially increased indoor capacities but also as a means of providing certain service to neighborhood outdoor users. The majority of the future data traffic demand, as discussed earlier, will originate from indoor locations and will be localized to certain geographical locations, mainly dense urban areas, and not the whole network. As demonstrated in the previous section, indoor-based small cells solutions serve as a key technology to provide indoor capacities with high speed data services, while the evolution of outdoor network elements for enhanced outdoor users is still unresolved. Considering the same traffic ratio (65–70 percent indoor and 30–35 percent outdoor), the outdoor traffic demand will also increase 300–350× (assuming 1000× increase in overall cellular data traffic). For outdoor service provisioning, due to relatively low traffic volume and high mobility users, mobile operators may still continue for a while to rely on the macrocellular layer to provide wide area coverage. This trend, however, will not last for long, as the recent advancements in wireless connectivity (e.g., for vehicles), supporting different applications ranging from infotainment and security to navigation and so on, will put stringent requirements on the mobile operators' infrastructure outdoors as well. Such innovations will demand high bit rates with anywhere anytime availability, which legacy outdoor deployments inherently lack. Traditional macrocellular deployments are only able to provide peak bit rates to relatively few users in a certain geographic location. This is attributed to the fact that due to large coverage areas of macrocells, users located near the cell site experience much better radio channel conditions and lower path loss than users at the cell edge, thus resulting in very uneven distribution of the achievable data rates throughout the cell coverage area. For next generation high-speed services, the distance between the eNode-B and UE has to be small enough to have minimum path loss and thus provide high SINR. Massive MIMO, with large antenna arrays, is one way to go [16]. However, due to its large size, it might not be feasible for the operators to deploy a large antenna array in urban downtown areas (e.g., where zoning restrictions apply).

Outdoor small cells can effectively address the inherent problem of macrocells. Due to their relatively small coverage footprint, the users are always within their vicinity, which helps reduce the path loss. Neverthless, the problem with outdoor small cells is that in order to have continuous coverage, small cells in sizeable numbers need to be placed quite close to each other, which poses problems for high mobility users. In the indoor environment, where the locations of indoor users rarely change, the indoor small cell solutions are able to provide better service. However, the situation in the outdoor environment is completely different as the majority of outdoor subscribers are highly mobile users. Deploying several hundred outdoor small cells to cover a certain downtown area can result in a large number of handovers, thus resulting in network signaling overload, which eventually will result in connection drops. Thus, in order to provide high-speed data services to high mobility users, especially in downtown areas where the majority of the capacity demand will be concentrated, operators will need to deploy solutions that offer both flavors of macro and small cells (i.e., deploying solutions that offer high-speed data services with a low degree of handovers). We believe that a possible solution lies in using *outdoor distributed antenna systems*, wherein distributed remote antenna nodes, usually deployed on public utility poles (e.g., streetlights) are clustered to form one large cell. Hence, through the distributed antenna nodes, if configured to transmit the same signal, the UE can experience highly stable signal strength throughout the DAS cell coverage area. On the other hand, if different nodes are configured to transmit independent signals, the amount of spatial reuse, and thus the local area capacity, is increased. Depending on the size of the cell, the number of nodes per DAS cell may vary from 5 to 100 or even more. Compared to a small cell, which typically has a cell range of 100–500 m, a DAS cell may cover one street block or a whole neighborhood. In the past, most DAS deployments were limited to indoor environments, usually in large malls, academic institutions, or commercial buildings with few outdoor installations. More recently, DAS deployments have started to make their way into the outdoor deployment arena, with the majority of the deployments taking place so far in North America. This is explored next.

For next generation high speed services, the distance between the eNode-B and UE has to be small enough to have minimum path loss and thus provide high SINR. Massive MIMO with large antenna arrays is one way to go.

Capacity Limitation of Traditional DAS Technology and the Dynamic DAS Concept

As mentioned in the previous section, the capacity demand outdoors is predicted to increase by 300–350× during the next few years, which operators need to fulfill. In a typical cellular network,

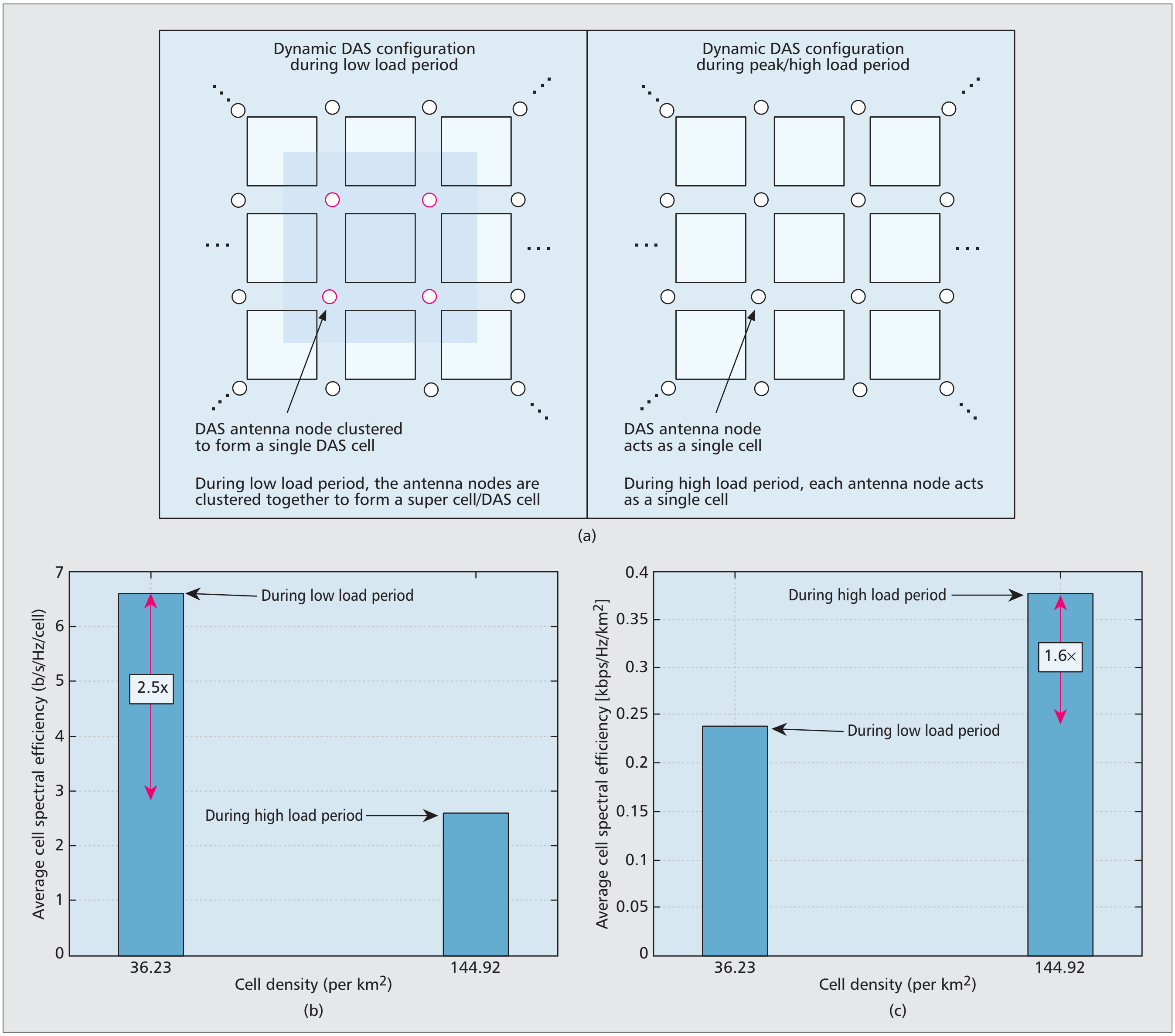


Figure 4. a) Dynamic distributed antenna system configurations during low network load time period and peak/high network load time period; b) average cell spectrum efficiency (b/s/Hz/cell); c) average network spectrum efficiency (kb/s/Hz/km^2), in different network load scenarios. The dotted lines in (a) indicate that the network is extending beyond what is shown.

the overall network traffic fluctuates according to time of day. An idle or low load is usually experienced, for example, during late night or early morning, while high/peak load is observed during the day and early evening. Also, the traffic pattern varies separately for outdoor (after office hours, on highways, boulevards, etc.) and indoor environments (during office hours and evenings). Hence, there is an inherent need to provide capacity dynamically outdoors whenever needed.

Traditional DAS systems work on static configurations (i.e., the clustering of the antenna nodes is preconfigured at the base station hub) and thus cannot efficiently deliver capacity in highly dynamic outdoor hotspot regions where the traffic varies geographically with time. Hence, most of the resources are wasted. In order to accomodate the 300–350× increase in outdoor traffic, the DAS network will need to dynamically reconfigure itself based on the instantaneous load, thus delivering the resources to parts of the network that experience high traffic load. The recent cloud RAN (C-RAN) technology aims to offer the capabilities, where hundreds of small cells or remote antenna units are connected via high-speed and low-latency backhaul links to a central hub site which houses the centralized processing servers [17], much like the traditional DAS system. In essence, the C-RAN is analogous to a dynamic DAS, which offers pooling of its resources at the central hub site and dynamically concentrates them to geographic areas with high capacity requirements (sport arenas, avenues, etc.). C-RAN technologies are among the hot topics being considered as key enablers for upcoming 5G networks [14]. In the following, we present an example scenario of a dynamic DAS deployment in an urban environment and evaluate its performance from the outdoor service provisioning point of view in terms of cell spectrum efficiency and network spectrum efficiency.

Cell and Network spectrum efficiency Performance of Dynamic DAS

Figure 4 shows the performance of a dynamic DAS network in two scenarios: during low network load times and high network load periods. A high network load in our example scenario is defined as a large number of users accessing the network at a certain time period in a certain location. When the number of users within the specific location increases, the network load rises. Hence, in order to accommodate more users within the specific geographic location, the DAS network, through a centralized processing server located in a central hub site, dynamically configures its remote antenna nodes to act as separate small cells. In this case, each antenna node is dedicated some resources from the central hub site, which houses the baseband transmission units. As such, the cell density within that geographic location increases, which translates into higher frequency reuse in that area, thereby accommodating more users. The tradeoff, however, in such a scenario is that due to closely spaced interfering sites, the spectrum efficiency inside individual cell deteriorates, resulting in lower cell-level throughput. In real networks, this cell-level performance reduction, stemming mostly from cell-edge users, can partially be compensated through intercell interference coordination (ICIC), and collaborative scheduling and/or beamforming methods. Subsequently, when the number of active users within that locality decreases, the network load goes down, and hence, the DAS network reconfigures itself by clustering the nodes in that region to form one super microcell (single DAS cell), as shown by the light blue area highlighted in Fig. 4a. The red colored dots are the nodes that form a single cluster of DAS cell nodes. The performance in terms of cell spectrum efficiency and network spectrum efficiency for the dynamic DAS network during peak and off-peak loads is illustrated in Figs. 4b and 4c, respectively. In this example, the network is deployed in an urban Manhattan grid environment, similar to the one depicted in Fig. 2b. The node (remote antenna) spacing is 130 m, which results in a node density of 144.92 nodes/km^2 (a continuous network is assumed). Moreover, in this example case, during low traffic load period, the network configures the nodes into a DAS configuration, wherein each DAS cell is composed of four remote antenna nodes, as shown in Fig. 4a. As such, the DAS cell density per square kilometer is 36.2 cells/km^2. During peak load period, the nodes are configured to act as individual small cells; hence, the cell density per square kilometer during peak time is 144.9 cells/km^2. Looking at the cell spectrum efficiency values, the gain in cell spectrum efficiency that can be achieved when the nodes are clustered into DAS cell configuration, is approximately 2.5 times more than in the peak load scenario, where the nodes operate as separate small cells. Hence, in low load scenarios, the dynamic DAS network can offer high-speed services to high mobility users with less number of handovers, due to the fact that the associated DAS nodes form one large cell. Coming then to the network spectrum efficiency performance during peak load, the gain in the network capacity, which is achievable through configuring each node as a small cell, is approximately 1.6 times more, which means that during peak load, the network is able to accommodate more users. From the handover point of view, during peak hours, the probability of high mobility users along the boulevards or high streets, where the majority of the data traffic is typically concentrated, is quite low due to crowded roads. Hence, the frequency of handovers during peak hours when the DAS nodes are acting as separate small cells, also reduces.

Conclusion

In this article, we have studied the performance of DenseNets from different deployment strategies' perspectives covering classical macro layer densification, the extremely dense indoor femto layer and outdoor dynamic distributed antenna system. The macrocell and ultra dense small cell deployment strategies have been evaluated from the cell spectrum efficiency, network spectrum efficiency, and network energy efficiency perspectives with an extreme level of densification and under full network load conditions to investigate and demonstrate the performance differences of these solutions when pushed to their capacity limits. The obtained results indicate that dedicated indoor solutions with densely deployed femtocells are much more spectrum-efficient and energy-efficient approaches to address the enormous indoor capacity demands compared to densifying the outdoor macro layer. Hence, we can conclude that to counter the growing concerns of the mobile operators related to the exponentially increasing amounts of mobile data toward the 5G era, an appealing solution is to deploy dedicated indoor solutions like femtocells, which offer a cost-effective and energy-efficient solution for indoor capacity demands. Also, from the indoor-to-outdoor service provisioning point of view, the mobile operators can partially leverage indoor-based femtocells to provide certain neighborhood coverage to low-mobility outdoor users, thereby offloading some of the traffic from the outdoor layer. This strategy can result in significant cost saving for mobile operators. Finally, from the outdoor service provisioning point of view, we have introduced and analyzed the dynamic outdoor DAS concept, which offers an efficient and capacity-adaptive solution to provide outdoor capacity on demand in urban areas by dynamically configuring the remote antenna units to either act as individual small cells or distributed nodes of a common central cell. One main purpose of this article is to raise awareness of the full network-level energy efficiency and spectrum efficiency potential of dedicated indoor systems, on one side, especially with increasing levels of wall penetration losses observed recently in modern buildings, and the reconfigurable capacity provisioning prospects of dynamic DAS solutions also closely connected to the emerging cloud-RAN concepts in the future.

We can conclude that to counter the growing concerns of the mobile operators related to the exponentially increasing amounts of mobile data toward the 5G era, an appealing solution is to deploy dedicated indoor solutions like femtocells which offer a cost-effective and energy-efficient solution for indoor capacity demands.

The obtained results indicate that dedicated indoor solutions with densely deployed femtocells are a much more spectrum-efficient and energy-efficient approach to address the enormous indoor capacity demands compared to densifying the outdoor macro layer.

Acknowledgments

The authors would like to extend their gratitude to the Finnish Funding Agency for Technology and Innovation, TEKES, for funding this research work as a part of the "Energy-Efficient Wireless Networks and Connectivity of Devices — Systems (EWINE-S)" project.

References

[1] Qualcomm Inc., "The 1000× Mobile Data Challenge," white paper, Nov. 2013.

[2] Ari Asp *et al.*, "Impact of Modern Construction Materials on Radio Signal Propagation: Practical Measurements and Network Planning Aspects," *Proc. IEEE VTC*, Seoul, Korea, May 2014.

[3] L. M. Correia *et al.*,"Challenges and Enabling Technologies for Energy-Aware Mobile Radio Networks," *IEEE Commun. Mag.*, vol. 48, Nov. 2010, pp. 66–72.

[4] S. F. Yunas *et al.*, "Impact of Macrocellular Network Densification on the Capacity, Energy and Cost-efficiency in Dense Urban Environment," *Int'l. J. Wireless & Mobile Networks*, vol. 5, no.5, 2013, pp. 99–118.

[5] S. F. Yunas *et al.*, "Deployment Strategies and Performance Analysis of Macrocell and Femtocell Networks in Suburban Environment with Modern Buildings," *Proc. IEEE LCN*, Edmonton, Canada, Sept. 2014, pp. 643–51.

[6] S. F. Yunas, *et al.*, "spectrum efficiency of Dynamic DAS with Extreme Downtilt Antenna Configuration," *Proc. IEEE PIMRC*, Washington, DC, Sept. 2014, pp. 2183–88.

[7] F. Richter and G. Fettweis, "Cellular Mobile Network Densification Utilizing Micro Base Stations," *Proc. IEEE ICC*, Cape Town, South Africa, May 2010, pp. 1–6.

[8] S. Tombaz, *et al.*, "Impact of Densification on Energy Efficiency in Wireless Access Networks," *Proc. IEEE GC Wkshps.*, Anaheim, CA, Dec. 2012, pp. 57–62.

[9] K. Hiltunen, "Total Power Consumption of Different Network Densification Alternatives," *Proc. IEEE PIMRC*, Sydney, NSW, Australia, Sept. 2011, pp. 1401–05.

[10] M. Polignano *et al.*,"The Inter-Cell Interference Dilemma in Dense Outdoor Small Cell Deployment," presented at *79th IEEE VTC*, Seoul, South Korea, May 2014.

[11] H. Ali-Ahmad *et al.*,"CROWD: An SDN Approach for DenseNets," *Proc. IEEE EWSDN*, Berlin, Germany, Oct. 2013, pp. 25–31 .

[12] Y. Lin and W. Yu, "Ergodic Capacity Analysis of Downlink Distributed Antenna Systems Using Stochastic Geometry," *Proc. IEEE ICC*, Budapest, Hungary, June 2013, pp. 3338–43.

[13] Naga Bhushan *et al.*, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 82–89.

[14] B. Bangerter *et al.*, "Networks and Devices for the 5G Era," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 90–96.

[15] Qualcomm Inc., "Neighborhood Small Cells for Hyper Dense Deployments: Taking HetNets to the Next Level," white paper, Feb. 2013.

[16] E. Larsson *et al.*,"Massive MIMO for Next Generation Wireless Systems," *IEEE Commun. Mag.*, vol. 52, Feb. 2014, pp. 186–95.

[17] China Mobile Research Institute, "C-RAN: The Road Towards Green RAN," white paper, v. 2.5, Oct. 2011.

Biographies

Syed Fahad Yunas (syed.yunas@tut.fi) received his B.Sc. degree in computer engineering from COMSATS Institute of Information Technology, Pakistan, and an M.Sc. degree in wireless communications from the University of Leeds, United Kingdom, in 2005 and 2007, respectively. Currently he is a postgraduate student working toward his D.Sc. (Tech.) degree at Tampere University of Technology (TUT), Finland. His research domain spans investigating novel cellular architecture for evolving mobile broadband networks and analyzing their techno-economic aspects.

Jarno Niemelä (jarno.niemela@elisa.fi) received his M.Sc. and D.Tech. degrees from TUT in 2003 and 2006, respectively. He has more than 12 years of experience in different radio network planning and optimization activities ranging from academia to industrial activities. He is currently a service manager at Elisa in the Department of Mobile Access Networks focused on managing the radio network planning process. Moreover, he is actively involved with different technology streams related to self-organizing networks, small cells, and indoor networks. Additionally, he acts as a part-time researcher at TUT concentrating on architectural aspects of future mobile radio networks.

Mikko Valkama (mikko.e.valkama@tut.fi) received his M.Sc. and Ph.D. degrees (both with honors) from TUT in 2000 and 2001, respectively. In 2003, he worked as a visiting researcher with the Communications Systems and Signal Processing Institute at San Diego State University, California. Currently, he is a full professor and Department Vice-Head with the Department of Electronics and Communications Engineering at TUT. His general research interests include communications signal processing, cognitive radio, full-duplex radio, radio localization, and 5G mobile cellular radio systems.

BEEcube
5G

Interference Coordination for Dense Wireless Networks

Beatriz Soret, Klaus I. Pedersen, Niels T. K. Jørgensen, and Víctor Fernández-López

ABSTRACT

The promise of ubiquitous and super-fast connectivity for the upcoming years will be in large part fulfilled by the addition of base stations and spectral aggregation. The resulting very dense networks (DenseNets) will face a number of technical challenges. Among others, the interference emerges as an old acquaintance with new significance. As a matter of fact, the interference conditions and the role of aggressor and victim depend to a large extent on the density and the scenario. To illustrate this, downlink interference statistics for different 3GPP simulation scenarios and a more irregular and dense deployment in Tokyo are compared. Evolution to DenseNets offers new opportunities for further development of downlink interference cooperation techniques. Various mechanisms in LTE and LTE-Advanced are revisited. Some techniques try to anticipate the future in a proactive way, whereas others simply react to an identified interference problem. As an example, we propose two algorithms to apply time domain and frequency domain small cell interference coordination in a DenseNet.

INTRODUCTION

Adding base stations has historically been the most important factor for increasing the capacity of cellular networks, and it is expected to persist in the upcoming years. Mobile operators are finding that very high traffic demands are typically concentrated in small geographical areas. To cope with this, small cells are the best match, since they can be opportunistically deployed in the hotspots, in a highly irregular way. Consequently, base station densification is going to be dominated by small cells. Besides that, taking new spectrum bands into use and techniques for efficient spectrum utilization will contribute to reach the challenging capacity targets. These very dense networks (DenseNets) can be seen as a natural evolution of today's Heterogeneous Networks (HetNets) [1, 2], inheriting most of their pros and cons.

However, DenseNets are also accompanied by a number of new challenges to be addressed. For example, backhaul will rise in importance [3]. With densification, the goal of operators is to deliver additional capacity and coverage with sufficient backhaul capacity and low latency without recurring operational expenditure (OPEX) charges, with solutions that range from fiber and Ethernet to wireless. Another important issue is mobility. Dense deployment of eNBs is challenging in a high-speed mobile environment, where frequent handovers may degrade the performance of the network. Numerous mobility enhancements and corresponding analyses have been studied in the context of HetNets [4, 5], and the investigations are expected to continue for DenseNets. The focus of this paper is the omnipresent interference, and how to combat it. Inter-cell interference is identified as the major limiting factor in Long Term Evolution (LTE) networks. Diverse interference management techniques have been included through successive releases of the LTE standard, from Rel. 8 to the latest completed Rel. 11. For example, solutions for interference coordination within the macro layer range from simple frequency domain methods [6] to more advanced coordinated multi-point (CoMP) techniques [7]. In the context of LTE HetNets, cross-tier interference (between the macro layer and the small cell layer) has been extensively investigated in the literature (e.g. [8]). With the anticipated small cell densification, the 3GPP work continues in Rel. 12 to have additional small cell enhancements [9], as well as coordinated multi-cell packet scheduling methods, referred to as enhanced CoMP.

The focus of this paper is on downlink interference, which becomes trickier in a dense deployment, with a more diffuse definition of aggressor cell and victim user. Here new techniques to deal with the co-tier interference are needed. In addition to network-based strategies relying on coordination among eNBs, advanced user equipments (UEs) will be equipped with interference cancellation capabilities that can further benefit from the knowledge about interfering transmissions under possible coordination by the network. Moreover, the mitigation techniques must be sufficiently dynamic to capture the variations of the interference, which can be very pronounced in a DenseNet where each cell serves a low number of users. For instance, we propose new time and frequency domain coordination strategies for dense clusters of small cells. The main idea is to have a proper resource division (time or frequency) by dynamically estimating the potential of the partitioning.

Beatriz Soret and Klaus I. Pedersen are with Nokia Networks.

Klaus I. Pedersen, Niels T.K. Jørgensen and Víctor Fernández-López are with Aalborg University.

0163-6804/15/$25.00 © 2015 IEEE

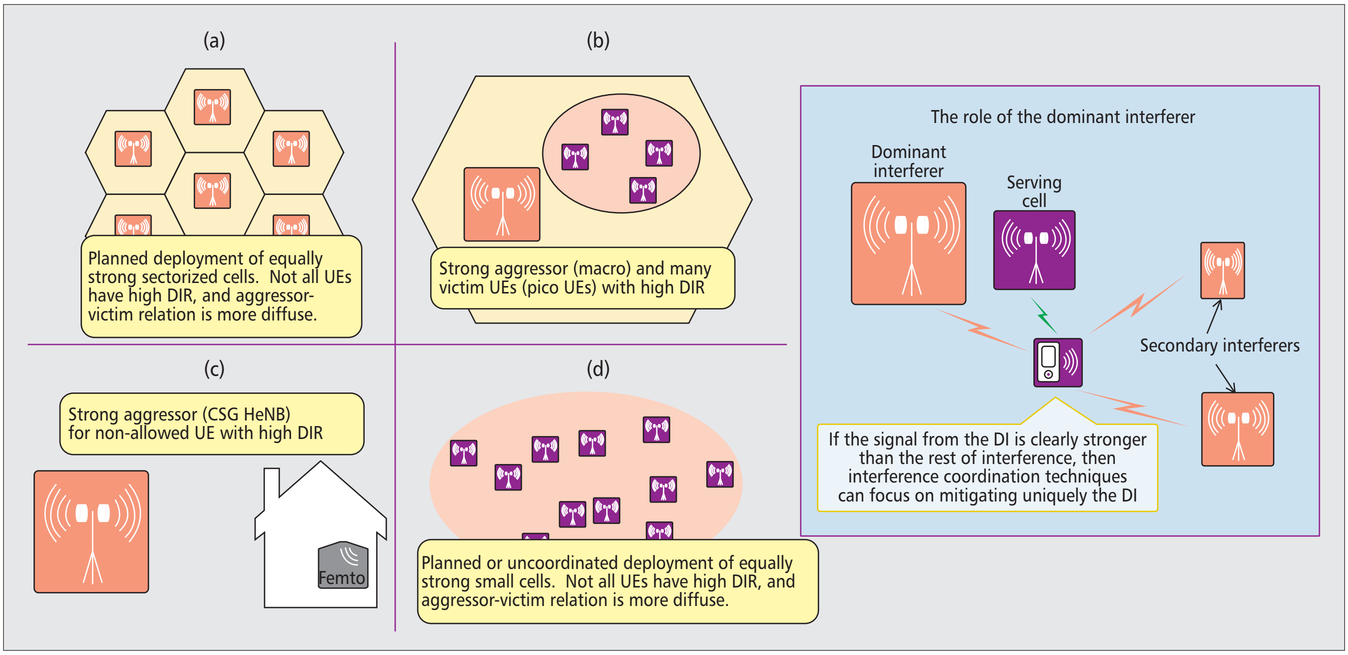

(a)
Planned deployment of equally strong sectorized cells. Not all UEs have high DIR, and aggressor-victim relation is more diffuse.
(b)
Strong aggressor (macro) and many victim UEs (pico UEs) with high DIR
(c)
Strong aggressor (CSG HeNB) for non-allowed UE with high DIR
Femto
(d)
Planned or uncoordinated deployment of equally strong small cells. Not all UEs have high DIR, and aggressor-victim relation is more diffuse.
The role of the dominant interferer
Dominant interferer
Serving cell
Secondary interferers
If the signal from the DI is clearly stronger than the rest of interference, then interference coordination techniques can focus on mitigating uniquely the DI

Figure 1. Interference scenarios and the role of the DI.

The rest of the paper is organized as follows. We first present the interference distribution in different 3GPP scenarios and a site-specific case in Tokyo, noting that the relation between aggressor and victim and the predominance of an interferer depend heavily on the particular scenario. Second, we give an overview of the available interference management methods. With more spread interference, there is still need for further development of inter-cell interference coordination (ICIC) techniques. We propose two solutions for the time and frequency domain small cell interference coordination, relying on either proactive or reactive schemes. In both cases, system level performance results are presented to demonstrate the benefits of small cell coordination in terms of higher end-user experienced throughput and lower outage probability. The article closes with concluding remarks.

Interference Scenarios and Statistics

Downlink interference can be mitigated from the network side by partially muting the interfering cells through a coordinated inter-cell algorithm. Another possibility is to let the UEs combat part of the interference by means of advanced receivers with interference cancellation (or suppression) capabilities. In any case, the choice of a proper interference management technique calls for a thorough study of the interference distribution between base stations and mobile users, where the interference sources for a UE are sorted from the strongest, the dominant interferer (DI), to the weakest. A good metric capturing the predominance of a single dominant interference is the dominant interference ratio (DIR), defined as the ratio between the DI and the rest of the perceived interference, shown mathematically as

$$DIR = \frac{I_{strongest}}{\sum\limits_{i \neq strongest} I_i + N} \qquad (1)$$

where $I_{strongest}$ is the power received from the DI, I_i is the power received from interferer i, and N is the thermal noise power. The improvement in signal to interference and noise ratio (SINR) from ideal interference cancellation of the DI is proportional to the DIR, giving a fine estimation of whether the strategy can focus uniquely on the DI, or if weaker interferers also need to be cancelled or suppressed. The quantities in Eq. 1 are time-variant, so the benefit from mitigating the DI is only fully achieved when conducted on a per-user basis and dynamic in time.

To illustrate the variation of the interference relations with the network topology, Fig. 1 draws four exemplary scenarios. Figure 1a is the traditional homogeneous network, deployed in a planned manner with equally strong sectorized macro cells, where not all the UEs perceive a high DIR and hence the aggressor-victim relation is diffuse. On the contrary, in a co-channel HetNet composed of macros and outdoor small cells sharing the same carrier (Fig. 1b), the macro cells are the clear aggressor for most of the small cell users, which are subject to strong downlink interference both in data transmission and control channels [8]. Another option is the deployment of indoor closed subscriber group (CSG) home eNB small cells (Fig. 1c), where macro users not belonging to the group cannot connect. Here, the small cell plays the role of aggressor to nearby indoor macro users that are not allowed to get service from the home eNB, resulting in so-called macro coverage holes. Also in this case, the definition of aggressor and victim is precise. Finally, small cells (indoor or outdoor) can be deployed on a dedicated carrier in a planned or unplanned manner (Fig. 1d), with

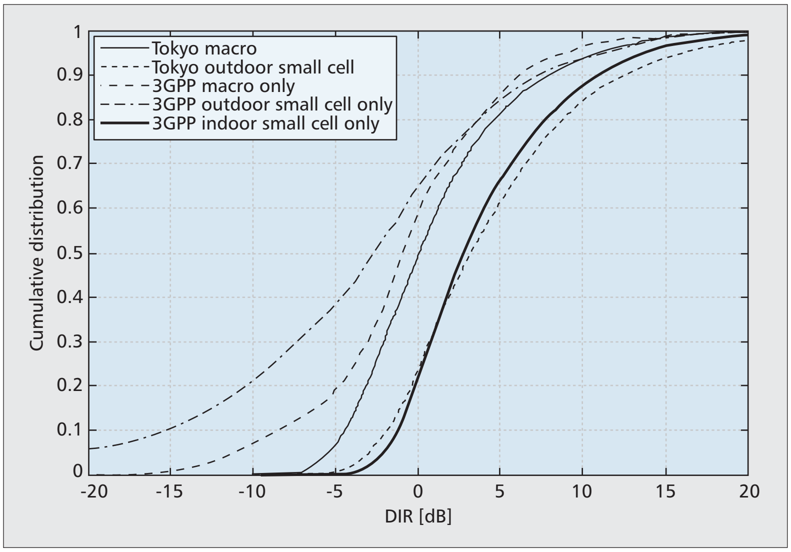


Figure 2. CDF of the DIR in different scenarios.

equally strong small cells and omni directional antennas. Similarly as for the homogeneous macro networks, not all UEs have a clear DI and the aggressor-victim relation is vaguer. Another factor is the potentially unplanned (and irregular) nature of this topology, which increases the probability of experiencing a high DIR.

To sum up, deployments of equally strong cells tend to experience a spread interference map (*co-tier interference*), where users do not necessarily perceive a clear aggressor or DI, but often multiple interfering signals of similar strength. The situation is exacerbated with densification: as the number of base stations per square meter increases, the chances of experiencing interference from more than one source also increase. On the other hand, interference between different layers (*cross-tier interference*) leads to higher values of DIR and has been widely investigated for HetNets. Finally, it is more likely to perceive a DI in more irregular deployments. With a high DIR, the benefit of applying some interference coordination or mitigation mechanism is obtained by focusing uniquely on the dominant interferer, while scenarios with low DIR are more challenging and need to deal with several interference sources.

In order to further illustrate the characteristics of different network deployments, Fig. 2 compares the empirical cumulative distribution function (CDF) of the DIR for various scenarios. Three generic 3GPP simulation scenarios as defined in [9] are considered, based on commonly accepted stochastic propagation models. The 3GPP macro-only deployment is composed of a regular grid of three-sector base stations deployed at 2 GHz, i.e. similar to the scenario in Fig. 1a. The 3GPP scenarios with clusters of small cells operate at 3.5 GHz. For the outdoor case, 10 small cells are randomly deployed in circular hotspot areas of 50 m radius. For the indoor case, a dual stripe multi-floor building block with one small cell per 100 m^2 apartment is assumed. In addition to the results from the standardized 3GPP cases, we also report results for a specific deployment in the city of Tokyo, Japan. Interference statistics are extracted for an area of approximately 1 km^2 around the Kinshicho Station. The buildings in this deployment area have an average height of 24 m and a maximum of 150 m. A total of 20 macro sites are deployed at 800 MHz (three-sector), 1700 MHz (three-sector), and 2100 MHz (six-sector), and at a height of 5 m above the building in its local area. The average macro inter-site distance equals 227 m (in contrast to the 500 m of the 3GPP case) with a standard deviation of 18 m. Moreover, 100 small cells are deployed at 3.5 GHz and at 5 m height in street canyons, placed mainly near the tallest buildings where the radio signal from the elevated macro-cells typically is weaker and more traffic can be offloaded. The statistics of the Tokyo case are separated for the macro and the small cell users.

Observing the curves in Fig. 2, the lowest DIR corresponds to the 3GPP outdoor small cell case, with dense clusters and a higher probability of coinciding with several active neighbors. On the other extreme, the highest DIR is observed for the Tokyo case, due to the more irregular and dense deployment, with the DIR of the 3GPP indoor case very close to the outdoor small cell layer in Tokyo. If a DIR of 3 dB, for example, is taken as a representative high value, less than 25 percent of UEs in the 3GPP outdoor small cell case will receive a clear benefit from mitigating the strongest interferer, whereas this percentage goes up to more than 50 percent in the small cell layer of the Tokyo scenario. The main learning here is that realistic dense networks (exemplified here by the data from Tokyo) may offer higher values of the DIR, and thus the gains of applying interference coordination might be higher as compared to the 3GPP scenarios. With lower values of the DIR, mechanisms mitigating the strongest interferer should be applied only for a selected subset of users.

Overview of Interference Mitigation Techniques

Extensive research related to LTE downlink interference mitigation has been performed in academia, industry, and standardization bodies such as 3GPP. Table I shows an overview of the different mechanisms. The interference problem can be addressed from the network side, the user side, or a joint action of both. Furthermore, some techniques try to anticipate the future in a proactive way, whereas others simply react to an identified interference source. The disadvantage of reactive solutions is that in highly dynamic environments the actions may happen too late. On the other hand, proactive approaches can lead to a waste of efforts and/or resources by trying to solve matters that may never materialize.

Within the network-based interference coordination category, the first group of solutions is based on resource partitioning, which can be conducted in the space domain, time domain, or frequency domain [9]. The simplest form of space domain resource partitioning is to use higher order sectorization in the macro site installations. As an example, upgrading from three-sector to six-sector macro sites is found to offer 50–80 percent capacity improvement depending on the spatial characteristics of the environment [10]. More advanced space domain techniques include coordinated beamforming and coordinated multi-point techniques [7].

The time and frequency domain resource partitioning techniques rely on blanking certain transmission resources in some cells to improve the perceived signal quality of those resources in the neighboring cells, resulting in a capacity loss for the cells blanking resources (called cost) and a benefit for the cells with reduced interference. The optimum blanking of resources can therefore be formulated as a value maximization problem, where the value (or the net benefit) equals the benefit minus the cost. The enhanced ICIC (eICIC) scheme is an example of time domain resource partitioning for co-channel macro and small cell deployments, where some transmission resources are blanked at the macro to improve the quality of the users served by the small cells [8]. The blanking is achieved by using the so-called "almost blank subframes" (ABS). Using ABS at the macro is found to offer promising performance improvements for co-channel macro and small cell cases, as the macro acts as an aggressor for many victim small cell users, and therefore the benefit can significantly exceed the cost.

Frequency domain resource partitioning can be realized by assigning different carriers to eNBs, or by using different OFDMA sub-carriers for transmission [6]. The simplest form is hard frequency reuse, where nearby eNBs use orthogonal frequency carriers. However, hard frequency reuse seldom results in the best performance for LTE. An alternative option is fractional frequency reuse (or soft frequency reuse), where some resources are reused by all eNBs, while others are dedicated to only certain eNBs. Furthermore, autonomous eNB mechanisms for dynamically choosing the best carrier(s) have been widely investigated in the context of femto cell networks [11]. In all cases, the potential of time and/or frequency domain inter-cell partitioning methods is fully exploited when they are dynamically adjusted in step with the time-variant behavior of the system and the traffic fluctuations. As examples of the former, [12] demonstrates the benefits of fast versus slow inter-cell coordination, while aspects of centralized versus distributed coordination are examined in [13].

Finally, the adjustment of the eNB transmit power is another network-based technique that has often been applied to closed subscriber group femto cells with the goal of reducing the cross-tier interference toward co-channel macro users [14].

An alternative to network-based interference coordination is to rely on advanced UE receivers with interference mitigation capabilities [3]. UEs with multiple antennas can exploit linear interference suppression techniques such as interference rejection combining (IRC). However, its applicability is limited. A UE equipped with M antennas has M degrees of freedom: one is used for the reception of its own stream; the remaining M-1 are available to exploit either diversity or interference suppression. For example, a UE equipped with two antennas and being served by an eNB using rank two has to use its single degree of freedom for inter-stream interference suppression. Yet the linear interference suppression at the UE can be boosted with network coordination. One example is to use rank coordination. The principle is to schedule victim UEs with rank one (single stream) on transmission resources where the neighboring cells also apply rank one transmission. By enforcing such inter-cell coordination, the highest gain from using IRC at the UE can be achieved. Similarly as for the resource partitioning techniques, the use of inter-cell rank coordination and IRC receivers presents a value that can be expressed as benefit minus cost. Here the benefit is the interference suppression gain offered by IRC, while the cost is the potential loss of throughput by restricting some cells to only use rank one transmission on certain resources.

The second variant of receiver-based interference mitigation is to apply non-linear interference cancellation, where the UE reconstructs the interfering signal(s) followed by subtraction before decoding the desired signal. These techniques are especially attractive for cancelling interference from semi-static signals such as common reference signals, broadcast channel, and synchronization channels, as already supported to a large extent in the latest LTE releases. However, applying non-linear interference cancellation to data channel transmissions is much more challenging, as the scheduling and link adaptation (i.e. selection of modulation and coding scheme) are highly dynamic, and conducted independently per cell. Hence, getting the most out of non-linear interference cancellation requires additional network assistance, and it is an ongoing work topic in 3GPP Rel-12 standardization [15]. The idea is to simplify the processing at the UE by providing a priori knowledge of the interfering signal characteristics such that the blind estimation of all their features can be reduced.

The network-based and receiver-based interference mitigation techniques in Table 1 essentially address the same problem: avoiding undesirable inter-cell interference. However, they have been typically treated separately in the literature. In principle, they are not mutually exclusive, but addressing the same problem independently from different perspectives can lead to some waste of effort. It remains to be further investigated how to maximize the synergies from both strategies. Thus, the new inter-cell interference challenges should be addressed by enforcing joint multi-cell cooperation techniques to fully exploit all degrees of freedom, as illustrated in Fig. 3. Further research on scheduler and link adaptation coordination between eNBs is required, providing additional a priori knowledge to UEs for interference cancellation, as well as exploiting recent advances in receiver signal processing techniques.

Further research on scheduler and link adaptation coordination between eNBs is required, providing additional a priori knowledge to UEs for interference cancellation, as well as exploiting recent advances in receiver signal processing techniques.

Small Cell Interference Coordination for DenseNets

Within the network-based ICIC category, we propose two methods to improve the performance of dense small cell networks: one proactive method using time domain ICIC, and a

Network based resource partitioning	Spatial-domain resource partitioning	Use of spatial filtering techniques. Simplest form is use of sectorized antennas. More advanced forms include use of arrays of transmit antennas or active antennas with coordinated beamforming between cells.
	Time-domain resource partitioning	Cells are time-synchronized and coordinate at which time-instances they transmit, such that there are time-instances where Cell A can serve its users without interference from Cell B. Also known as coordinated muting. Examples include 3GPP defined techniques such as eICIC and CoMP.
	Frequency-domain resource partitioning	Include options such as using hard or soft frequency reuse between neighboring cells. The frequency-domain resource partitioning can be on PRB resolution, or on carrier resolution if having networks with multiple carriers. The latter is also referred to as carrier-based ICIC.
Network based transmit power control	Transmit power control per cell	Adjustment of transmit power per cell to improve the interference conditions. Examples include 3GPP defined techniques for femto cell transmit power calibration to reduce interference toward co-channel macro-users.
UE based interference mitigation	Interference suppression	Interference suppression by means of linear combining of received signals at the UEs antennas. Examples of such techniques are optimal combining and interference rejection combining.
	Interference cancellation	Interference cancellation with non-linear techniques where the UE estimates one or multiple interfering signals and subtracts them from the received signal, followed by detection of the desired signal. Examples include successive or parallel interference cancellation schemes.
	Network assisted interference mitigation	Schemes where the UE receives additional assistance information from the network to facilitate more efficient interference mitigation. This includes cases where the UE receives a priori information of interfering signals that it should suppress. The simplest example is common reference signal (CRS) interference cancellation (IC), where the UE receives information related to neighboring cell CRS characteristics to enable easier non-linear IC of those.
Joint network and UE based nterference mitigation	Exploiting all degrees of freedom for maximizing the system performance	Hybrid schemes with joint multi-cell coordination to maximize the benefits of both network based and UE receiver based interference mitigation techniques. One example is to use inter-cell rank coordination, such that a UE with the capability of linear interference suppression of a few strong interfering streams is primarily scheduled when its serving and interfering cells are transmitting with rank-1.

Table 1. Overview of downlink interference mitigation toolbox.

second reactive scheme that relies on carrier domain ICIC. The time domain algorithm is applied to clusters of outdoor small cells, whereas the carrier domain solution has been evaluated for indoor deployments. In both cases it is required that the algorithm adapt to changing traffic conditions, created by a dynamic birth-death traffic model with a fixed payload per call. When the payload has been successfully delivered, the call is terminated.

Proactive Time Domain ICIC

In the time domain, some subframes are muted in the small cell layer in order to mitigate the interference to the victim users. As seen in the statistics in Fig. 2, the definition of aggressor and victim is not straightforward in dense clusters of cells, and deciding which small cell to mute and when to do it is not trivial. Even within the same cell, users perceive different neighbor small cells as their main aggressor.

The muting actions are only taken if a small cell is identified as an aggressor. Otherwise, normal transmission is used. Therefore, a key aspect of the algorithm is the identification of victim users and their aggressors. With the goal of improving the coverage user throughput, defined as the 5th percentile user throughput, without compromising the average user throughput, the identification of a victim user is twofold. First, the ratio between the received signal from the serving cell and the DI has to be below a threshold (set to 10 dB in the simulations). Second, the DIR has to be above 3 dB (the DI to be perceived at least at double power as compared to the rest of interference). If both conditions are met, the user is classified as a victim user and its

DI is identified as the aggressor cell, which will be requested to mute. The muting action is reverted when the victim user that triggered a muting leaves the system.

The muting coordination among small cells is especially challenging when one small cell is simultaneously aggressor and serving a victim user. In these cases it is necessary to coordinate the muting actions among small cells to avoid situations in which the cell serving the victim user is muting at the same time as the aggressor. This coordination is attained with a proactive approach, in such a way that each cell has some pre-assigned "good" time slots with improved SINR conditions and some "bad" slots where it may be asked to mute. The pattern of these pre-assigned time-slots is set a priori. As the densification grows, it is not convenient to apply the algorithm at a full cluster level as it leads to too complex coordination, and instead coordination within subclusters of small cells is recommended. The small cell subcluster division can be done based on the past history UE measurements and/or small cell network listening mode (NLM) measurements to identify interfering cells that should belong to the same subcluster [13].

Reactive Carrier Domain ICIC

As a second example of network-based interference coordination, we present a reactive carrier domain ICIC solution. The goal is to orchestrate a proper use of the component carriers (CC) to have all users served with at least a certain minimum data rate, expressed by the guaranteed bit rate (GBR). By default, all the small cells utilize all the available CCs (reuse 1 strategy).

The identification of victim users experiencing too low service rates, i.e. below the promised GBR, is the criterion to trigger the reactive actions. If the small cell serving the victim user is not using all its CCs, it can choose to enable more CCs to increase the available bandwidth. It can also choose to request interfering small cells to stop using certain CCs to reduce the experienced interference at the victim user. For each of the possible hypotheses to improve the performance of the victim user, the corresponding value (benefit minus cost) is estimated, followed by taking the action that results in the highest positive value. As an example, the hypothesis corresponding to taking more CCs into use for the small cell serving the victim user will result in a benefit for that cell, but also a potential cost in the neighboring cells that will experience increased interference. Similarly, if a CC is switched off in cell A it will result in a performance loss (cost) for users served by cell A, while users experiencing interference from cell A will experience less interference (benefit). For the sake of simplicity, not all the possible hypotheses are evaluated, but only those that involve neighboring cells acting as a DI for the identified victim user. Finally, when a user that has previously triggered the carrier domain ICIC framework leaves the system, the prior actions aiming at improving the performance for that user can be reverted.

It is worth noting that the benefit and cost calculations require information to be shared between the small cells over the backhaul. However, the information is rather limited, and is not considered sensitive to typical backhaul latencies of 10-50 ms.

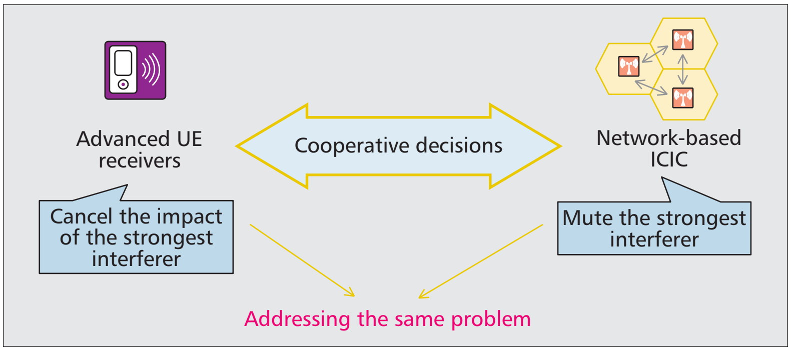

Advanced UE receivers
Cooperative decisions
Network-based ICIC
Cancel the impact of the strongest interferer
Mute the strongest interferer
Addressing the same problem

Figure 3. Cooperative network-based and UE receiver-based interference mitigation.

Performance Gains

A network layout following the guidelines in [9] is simulated. The considered 3GPP Rel-12 small cell scenarios with clustered outdoor cells and indoor cells are in line with the descriptions given for Fig. 2. For the case with outdoor clusters, we consider an ultra dense case with 12 small cells per cluster, whereas indoor small cells are in a dual stripe building block. The system-level simulator follows the LTE specifications, including detailed modeling of major radio resource management functionalities such as packet scheduling, hybrid automatic repeat request (HARQ), link adaptation, 2×2 closed loop single-user MIMO with dynamic precoding, and rank adaptation. Proportional fair (PF) scheduling is applied independently at each cell. The finite payload per user is 0.5 Mbytes. For the simulations of outdoor small cell clusters, we use an open loop traffic model with Poisson call arrivals and an average offered load per cluster area ranging from 50 Mbps to 110 Mbps. The simulations for the indoor small cell cases assume a closed-loop traffic model with a constant number of users per building block, with a new call generated immediately after an existing call is completed.

In Fig. 4a the user throughput gain of time domain ICIC is presented as a function of the offered load. As expected, the relative gain increases with the offered load of the system, both in 5 percentile and 50 percentile user throughput, going up to 40 percent and 25 percent respectively for the highest simulated load. On the other hand, no significant gains were observed for values of offered load below 50 Mbps. This makes good sense: at low load, few users are active at the same time, and the probability of experiencing strong interference from a neighbor small cell decreases. In Fig. 4b the maximum muting ratio (corresponding to the small cell muting a larger percentage of time in the simulation) and the average muting is plotted, as a function of the offered load of the system. As the offered load increases, the percentage of muting in the system also increases, since the condition triggering the muting actions is met more often.

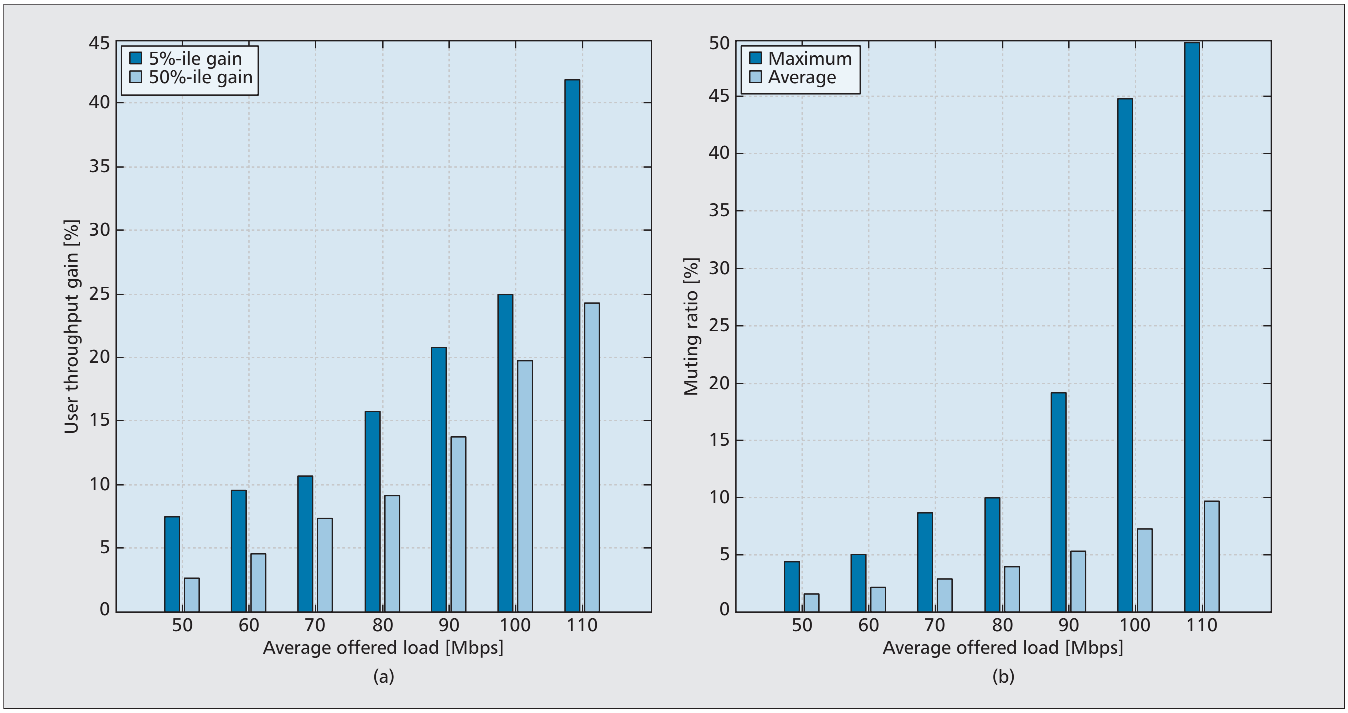


Figure 4. Time domain ICIC with outdoor small cell DenseNets: performance results: a) user throughput gain vs. average offered load; b) maximum and average muting ratio vs. average offered load.

In Fig. 5 the performance of carrier domain ICIC with four CC per small cell is shown. Figure 5a shows the outage probability of having users experiencing a service rate below their GBR versus the offered load (expressed by the average number of users per small cell). Results are reported for both the plain frequency reuse case (without any interference management) and for the proposed reactive carrier domain ICIC scheme. Similarly to the results of the time domain ICIC, there is no gain from applying interference coordination at low load with only a few users per small cell. As expected, the improvement in outage becomes significant as the load increases, allowing one more user per small cell when the carrier domain ICIC is enabled. Indeed, the increase in capacity goes up to 25 percent: four users with reuse one versus the five users of carrier domain ICIC. The probability mass function for the number of used CCs per small cell is reported in Fig. 5b for each offered load. With only one user per small cell on average, it is observed that 94 percent of the cells use all four CCs, i.e. the carrier domain ICIC is seldom triggered. As the load increases, the interference coordination is applied more often, with only 18 percent probability of using all four CCs per small cell.

Conclusions

In this paper we have discussed the role of the interference for a variety of deployments, ranging from homogeneous macro-only networks to dense small cell networks. The first step has been motivating the terminology of aggressor and victim and the dominant interference ratio (DIR), as effective elements for investigating the advisability of interference coordination. Interference statistics for generic 3GPP simulation scenarios and a site-specific case in Tokyo are compared, showing a larger potential of applying interference management techniques in the latter case. An overview of the huge variety of interference management techniques is also presented, and the best solution for a given network will depend on factors such as the deployment, the desired optimization goal, or the UE capabilities. Hybrid schemes of network-based interference coordination and user-based interference suppression by means of advanced receiver signal processing are identified as an area that requires further research. Finally, we have proposed two algorithms to apply either proactive time-domain or reactive carrier-domain co-tier interference coordination with different optimization goals. The main idea is to have a proper resource division (time or frequency) by dynamically estimating the potential of the partitioning. The performance results show gains of 25–40 percent user throughput and 25 percent in capacity. In conclusion, we have essentially shown that evolution to DenseNets opens new opportunities for interference coordination research.

Acknowledgement

The authors want to thank Laura Luque from Aalborg University for her help with the site-specific scenario.

References

[1] A. Ghosh *et al.*, "Heterogeneous Cellular Networks: From Theory to Practice," *IEEE Commun. Mag.*, vol. 50, no. 6, June 2012, pp. 54–64.

[2] J.G. Andrews, "Seven Ways that HetNets are a Cellular Paradigm Shift," *IEEE Commun. Mag.*, vol. 51, no. 3, Mar. 2013, pp. 136–44.

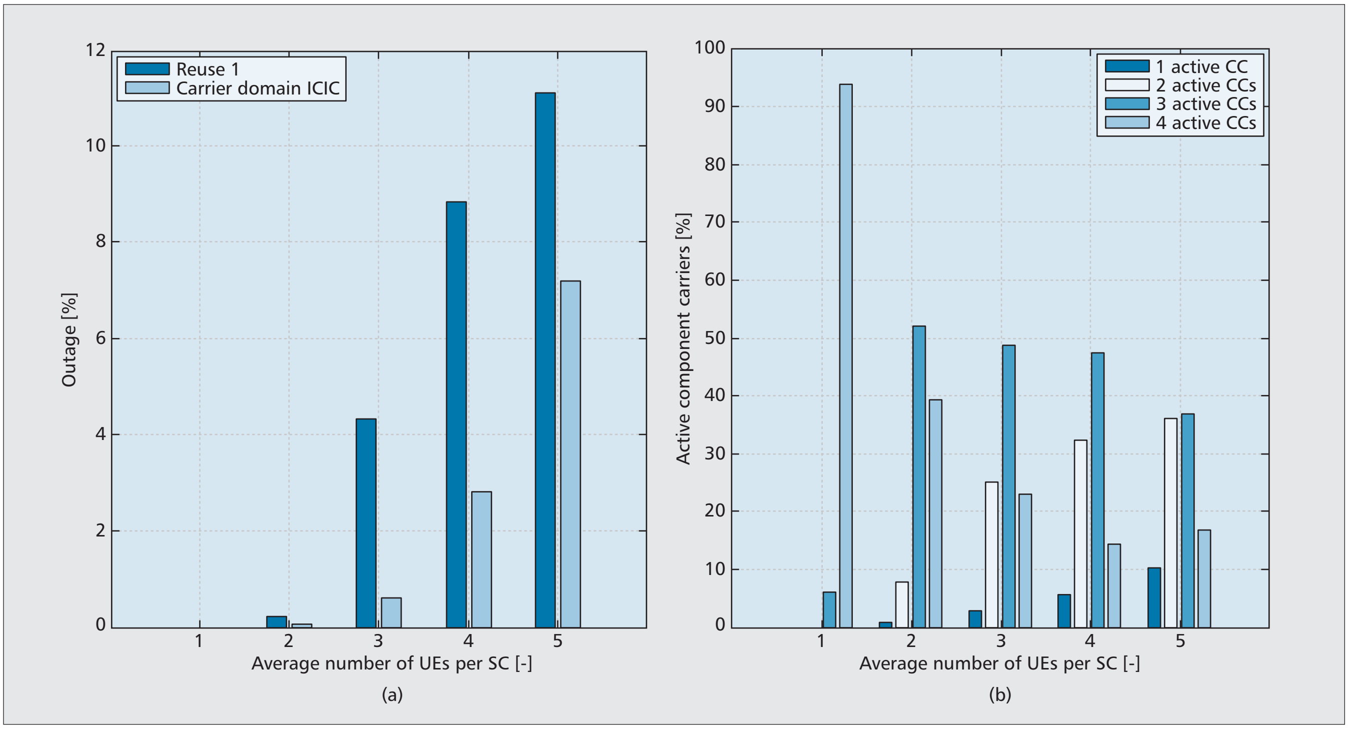


Figure 5. Carrier domain ICIC with indoor small cell DenseNets: performance results: a) outage vs. average number of UEs per small cell; b) probability mass function for the number of used CCs vs. average number of UEs per small cell.

[3] N. Bhushan *et al.*, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 82–89.

[4] S. Barbera *et al.*, "Improved Mobility Performance in LTE Co-Channel HetNets Through Speed Differentiated Enhancements," *IEEE Proc. Globecom, Wksp. Heterogeneous, Multi-hop, Wireless, and Mobile Networks*, Dec. 2012, pp. 426–30.

[5] A. Prasad *et al.*, "Energy-Efficient Inter-Frequency Small Cell Discovery Techniques for LTE-Advanced Heterogeneous Network Deployments," *IEEE Commun. Mag.*, vol. 51, no. 5, May 2013, pp. 72–81.

[6] G. Boudreau *et al.*, "Interference Coordination and Cancellation for 4G Networks," *IEEE Commun. Mag.*, vol. 47, no. 4, Apr. 2009, pp. 74–81.

[7] J. Lee *et al.*, "Coordinated Multipoint Transmission and Reception in LTE-Advanced Systems," *IEEE Commun. Mag.*, vol. 50, no. 11, Nov. 2012, pp. 44–50.

[8] K. I. Pedersen *et al.*, "Enhanced Inter-Cell Interference Coordination in Co-Channel Multilayer LTE-Advanced Networks," *IEEE Wireless Commun. Mag.*, vol. 20, no. 3, June 2013, pp. 120–27.

[9] 3GPP, TR 36.872 "Small Cell Enhancements for E-UTRA and E-UTRAN — Physical Layer Aspects," v. 12.0.0, Sept. 2013.

[10] S. Kumar *et al.*, "Performance Evaluation of 6-Sector-Site Deployment for Downlink UTRAN Long Term Evolution," *IEEE Proc. Vehic. Tech. Conf.*, VTC 2008-Fall, Sept. 2008.

[11] L. Garcia *et al.*, "Autonomous Component Carrier Selection for 4G Femtocells — A Fresh Look at an Old Problem," *IEEE JSAC*, vol. 30, no. 3, Apr. 2012, pp. 525–37.

[12] B. Soret *et al.*, "Fast Muting Adaptation for LTE-A HetNets with Remote Radio Heads," *IEEE Proc. Globecom*, December 2013.

[13] R. Agrawal *et al.*, "Centralized and Decentralized Coordinated Scheduling with Muting," *IEEE Proc. Vehic. Tech. Conf.*, May 2014.

[14] D. López-Pérez *et al.*, "OFDMA Femtocells: A Roadmap on Interference Avoidance," *IEEE Commun. Mag.*, vol. 47, no. 9, Sept. 2009, pp. 41–48.

[15] 3GPP, TR 36.866 "Study on Network-Assisted Interference Cancellation and Suppression (NAICS) for LTE," v. 12.0.0, Mar. 2014.

BIOGRAPHIES

BEATRIZ SORET (beatriz.soret@nsn.com) received her M.Sc. and Ph.D. degrees in 2002 and 2010, respectively, from Málaga University, Spain. She is currently a radio research specialist at Nokia Networks in Aalborg, Denmark. Her main research interests are related to radio resource management, heterogeneous networks, and quality of service for 4G and 5G networks.

KLAUS I. PEDERSEN (klaus.pedersen@nsn.com) received his M.Sc. E.E. and Ph.D. degrees in 1996 and 2000, respectively, from Aalborg University, Denmark. He is currently with Nokia Networks in Aalborg, where he is a senior wireless network specialist. His current work is related to radio resource management and 3GPP standardization of LTE. He has been appointed as apart-time professor at Aalborg University in the Wireless Communications Networks (WCN) section.

NIELS T. K. JØRGENSEN (nj@es.aau.dk) received his BSE and MSE from Aalborg University, Denmark in 2007 and 2009, respectively. Since 2011 he has been pursuing his Ph.D. degree, also at Aalborg University. His research interest focuses mainly on radio resource management and interference management in heterogeneous networks or small cell networks.

VÍCTOR FERNÁNDEZ-LÓPEZ (vfl@es.aau.dk) received his Master's degree from the University of Granada, Spain, in 2012. He is presently a doctoral student in the Wireless Communications Networks section, Aalborg University, Denmark. His primary research interests are LTE resource management and heterogeneous networks.

Understanding Channel Selection Dynamics in Dense Wi-Fi Networks

Akash Baid and Dipankar Raychaudhuri

ABSTRACT

This paper aims to explain and analyze a growing problem in dense-urban wireless networks, that of co-existence between low-cost residential access points (APs) and actively-managed service provider APs in overlapping spatial, frequency, and time domains. Through detailed simulations and testbed experiments, the impact of increasing density of highly-adaptive service provider APs on the performance of typical residential APs is measured in terms of their respective channel assignment schemes. Simulation results with dense deployment of up to 500 APs/sq.km. show the benefits of centralized channel assignments, even in the presence of independent APs. In addition, it is shown that for a fixed AP density, an increase in the percentage of AP under the centralized scheme results in an increase in the throughput of surrounding independent APs. The broader implications of the simulation findings are discussed in order to develop a better macro-level understanding of dense Wi-Fi networks.

INTRODUCTION

Since their introduction in the early 2000s, the deployment of wireless local area networks (WLANs) has been constantly growing. While most urban residences and offices now already use WLANs, the recent spurt in WLAN density has been due to large-scale deployments by mobile operators and broadband Internet providers, collectively termed as service providers (SP) Wi-Fi [1]. This rapid rise in the number of Wi-Fi access points (APs) has led to an interesting mix of deployments where residential, enterprise, and SP APs operate on the same spectrum (and can thus interfere with each other), but enterprise and SP APs are actively managed and can usually adapt to interference much better than residential APs due to better and more expensive hardware and software. Figure 1 shows the combined percentage of enterprise and SP access points (of the total APs observed) in a representative 1 sq. km. area of four major US cities, as per the crowd-sourced WiGLE.net database [2]. A clear trend of an increasing percentage of "managed WLANs" can be observed, especially since the beginning of 2012.[1]

Since the coverage regions of the enterprise/SP APs often overlap with that of residential APs, this growth in actively managed APs can result in performance problems for the residential APs. An immediate example of the potential problem is the disparity between the channel selection schemes used in residential and managed WLANs. Most low-cost residential APs either operate on a fixed channel or change channels only upon power cycle, while most enterprise and SP APs incorporate centralized, adaptive channel assignment schemes. Thus in areas where both types of APs are present, the residential APs can potentially be cornered into higher interference channels, while the managed APs adapt their channels in response to interference. In this paper we target such mixed-deployment problems and build the understanding toward the key question: *What is the impact of the increasing density of highly-adaptive enterprise/service provider APs on the performance of typical residential APs and vice versa?*

In order to measure the performance of different types of APs in extremely dense networks, we extend Liew's Maximum Independent Set (MIS) model for channel share estimation [4]. In their seminal paper (which has since laid the foundation for throughput-optimal CSMA [5]), Liew *et al.* proposed an approximate but highly-accurate technique to calculate the channel share of an AP given the contention graph of the nodes surrounding that AP. Using the original MIS model for dense network graphs comprising hundreds of nodes runs into computational bottlenecks since the process involves finding all maximum independent sets of a graph, a classical NP-hard problem with a long standing bound of exponential complexity [6]. As such, in this paper we propose an approximation mechanism to parametrize the balance between computational complexity and desired accuracy.

Using this approximation technique, we measure the performance of a typical centralized channel assignment algorithm in the presence of a varying number of residential APs through dense-deployment simulations. The simulation scenarios are designed to reflect the current deployment mix in urban areas (5–25 percent managed and the rest residential), and also the possible continuation of the trends shown in Fig. 1, for example 50–75 percent managed APs. Different channel assignment schemes are assumed for the low-cost residential APs, in particular, static default, random, and least congested channel schemes. A key finding from the

Akash Baid and Dipankar Raychaudhuri are with Rutgers University.

Research supported by NSF EARS grant CNS-1247764.

[1] The percentage share of managed WLANs is calculated by matching the first three bytes of the logged MAC addresses with the IEEE OUI record of the top ten enterprise/SP WLAN vendors. Since there are other enterprise/service-provider WLAN vendors, the percentage share estimated here is a lower bound. See [3] for further details.

0163-6804/15/$25.00 © 2015 IEEE

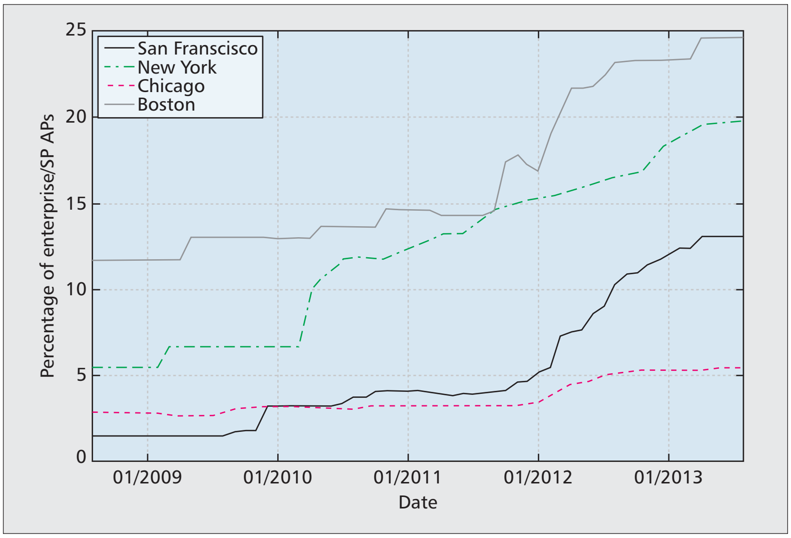


Figure 1. Percentage of APs from enterprise/SP WLAN vendors out of all observed APs from the WiGLE.net database [2].

simulations is that, while the trend of an increasing percentage of managed APs would improve the overall utilization of the ISM band, at high densities the existing managed APs would perform worse and the existing residential APs would perform better. The intuition behind this result is that, to an extent, the better performance of the managed APs over residential APs is because of the non-optimal choices made by the latter; and as more and more APs improve their resource-usage choices, the potential gains for managed APs is reduced due to the overall capacity of the spectrum being bounded.

The key contributions of this paper are:

- We propose a parametric approximation scheme to extend known channel estimation models for extremely dense network graphs. The key parameter in the model controls the tradeoff between accuracy and computation time.
- We study the performance of Wi-Fi APs in homogeneous as well as mixed settings, i.e. a fraction of APs in a region use simple static channels, while others are managed by a central controller.
- We highlight the issue of inverse correlation between the percentage of managed APs and their performance relative to non-managed APs through both simulation and experimental results.
- We discuss the broader implications of the observed results in terms of the performance of unlicensed band nodes in dense settings.

Modeling AP Channel Share in Dense Deployments

The number of available channels in Wi-Fi is substantially less than what is required to build a conflict-free graph in dense settings. Hence all practical channel assignment schemes must assign the same channel to multiple APs in range of each other. A channel assignment scheme working with k available channels converts the distance-based graph, i.e. one in which an edge exists between two nodes if they are in carrier sense range of each other irrespective of the operating channel, to k derived-graphs. A node appears in derived-graph i if it has been assigned channel i, and a link in the original distance-based graph is transferred to the derived-graph i only if both its end-points are in i. Given such derived-graphs, a general model for the channel share of each AP as per the underlying CSMA protocol has proven to be extremely elusive, except for the case of a completely connected graph for which Bianchi's work provides an accurate model [7]. For a completely connected graph with N nodes, the channel share of each node comes to approximately $1/N$.

Liew's MIS Model

Liew *et al.* [4] proposed the following simple technique to calculate the approximate channel share of each node. Given a contention graph, first calculate its maximum independent sets (MISs). An independent set is a set of vertices, no two of which are connected by a link in the graph, and the maximum independent sets are such sets with the highest number of elements. The normalized throughput of each node in the graph is then given by the ratio of the number of MISs that node appears in to the total number of MISs. Performance results from experiments done with physical Wi-Fi devices have been shown to closely match the estimates given by this MIS model [4].

While being derived from a theoretical analysis of the underlying CSMA networks, the key intuition behind the accuracy of the MIS model is that among the 2^N possible states comprised of each node of a N node graph being on or off, the CSMA protocol largely favors the "greedy" states, i.e. the states that result in the maximum number of nodes transmitting simultaneously. Further, all such greedy states are equally probable, and thus the throughput of each node is dependent on how many greedy states it appears in, relative to the total number of such states.

Parametric Approximation of the MIS Model

Although simple to reason about, the problem with utilizing this MIS model is that computing all maximum independent sets of a graph is a classical NP-hard problem with a longstanding bound of exponential complexity [6]. As such, we propose the following approximation mechanism to parametrize the balance between computational complexity and desired accuracy.

Since computing the MISs of the complete graph is computationally expensive, we use the same MIS model per node over a neighborhood-graph centered around each node. The neighborhood-graph is defined by a parameter termed *span* which can range from 0 to the diameter of the graph. For a selected span s, the neighborhood-graph of a node i is formed of all the nodes at a graph-distance of less than or equal to s. For each node j at a distance exactly equal to s from node i, all directly connected nodes that are not already included in the neighborhood-

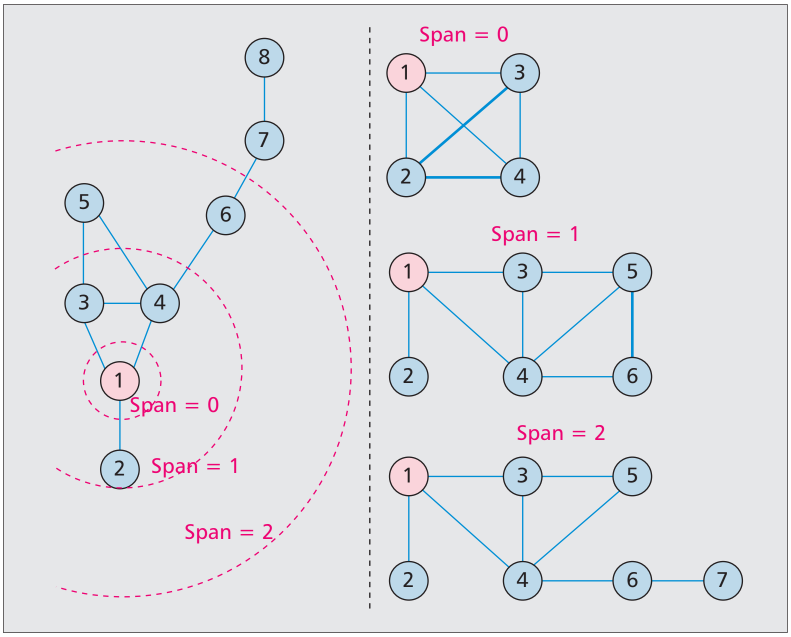


Figure 2. Formation of the neighborhood-graph for approximating the MIS model.

graph of *i* are added to it but the connectivity between such nodes is assumed to be a clique. The process is illustrated in Fig. 2, which shows the process for building neighborhood-graphs of different spans around node 1. Note that for the span 0 graph, nodes 2, 3, and 4 are included but links 2–3 and 2–4 are added even though they are not present in the original graph.

The intuition behind the step of clique-formation at the edge of the span is to invoke the standard 1/*N* model beyond the point of the neighborhood-graph. This results in the computed channel share to be exactly equal to that found through the 1/*N* model for span 0 and equal to that derived from the MIS model for maximum span. Figures 3a and 3b show the mean error compared to maximum span, and the time required for computation respectively when varying the span from 0 to 2 and the number of nodes in the graph from 20 to 50. All values are averaged over 100 random initiations of the graph. As is clear from these plots, while computing the span 0, i.e. the 1/*N* model is extremely fast, it can result in large errors; increasing the span decreases the error but results in a corresponding increase in the computation time. For a given application, the value of the span parameter should be chosen according to the requirements of accuracy and computation time. For all the simulation results presented next, we use a value of *span* = 2 since for the size of the graphs considered, a higher span results in prohibitively larger computation times.

Analyzing Channel Assignments in Mixed Deployments

The approximate MIS model defined earlier provides a scalable mechanism to estimate the saturation throughput of APs given the deployment topology and the channel assignment. In this section, we use that model to study the performance of different channel assignment mechanisms under different assumptions about the mix of residential vs. enterprise/hotspot APs.

Simulation Description

All results presented in this section are based on MATLAB simulations of dense AP deployments in a 1 sq. km. area. To exactly model the performance perceived by clients in a realistic deployment, the simulation must consider, at the least:

- Environment-dependent pathloss, shadowing, and multipath, including wall losses.
- The number, placement, and capabilities of client devices.
- The offered load and its variation for each client.
- The policy of the AP for scheduling multiple backlogged clients (note that this is not specified by the 802.11 standard).
- Capture effect, based on relative signal strength and timing of interfering signals.

Accounting for all these factors can make the simulations extremely intractable, especially when simulating extremely dense deployments. As such, we consider a much simplified simulation setting that retains the qualitative nature of the tradeoffs involved but admittedly misses some of the finer nuances involved in wireless communications.

In order to focus on node-starvation and similar network-level effects, we limit the granularity of simulations to APs, i.e. measuring the throughput achieved at each AP instead of each client. This relieves us from the task of modeling AP load-distribution policy, client locations, and capabilities. We assume a downlink saturation scenario, which translates to the assumption of each AP always having one or more connected clients whose data demand is enough to prevent the AP from being idle when it gets access to the channel. We consider a purely distance-based interference model. If two APs are within carrier sense range (assumed to be 100 meters), there exists a link between them in the contention graph. Each channel assignment scheme is assumed to be working with three orthogonal channels, as in the 2.4 GHz band. Similar results can be obtained for a regime with more channels or with non-orthogonal channels by considering a channel overlap dependent sharing model [8].

Metrics: Normalized Throughput and Starved Nodes — We use two key performance metrics throughout this study. The first metric is mean normalized throughput received by an AP. As mentioned above, the throughput "received by an AP" reflects the combined throughput that all clients connected to the AP would be expected to receive. The term normalized is used to indicate that all throughputs are expressed as a fraction between 0 and 1; an AP with a normalized throughput of 1 gets access to the channel 100 percent of the time, whereas a normalized throughput of 0 indicates that the AP never sees the channel free for transmission. This is calculated using the approximate MIS model described earlier. The other metric we focus on is the percentage of starved nodes, as estimated from the MIS model. A channel

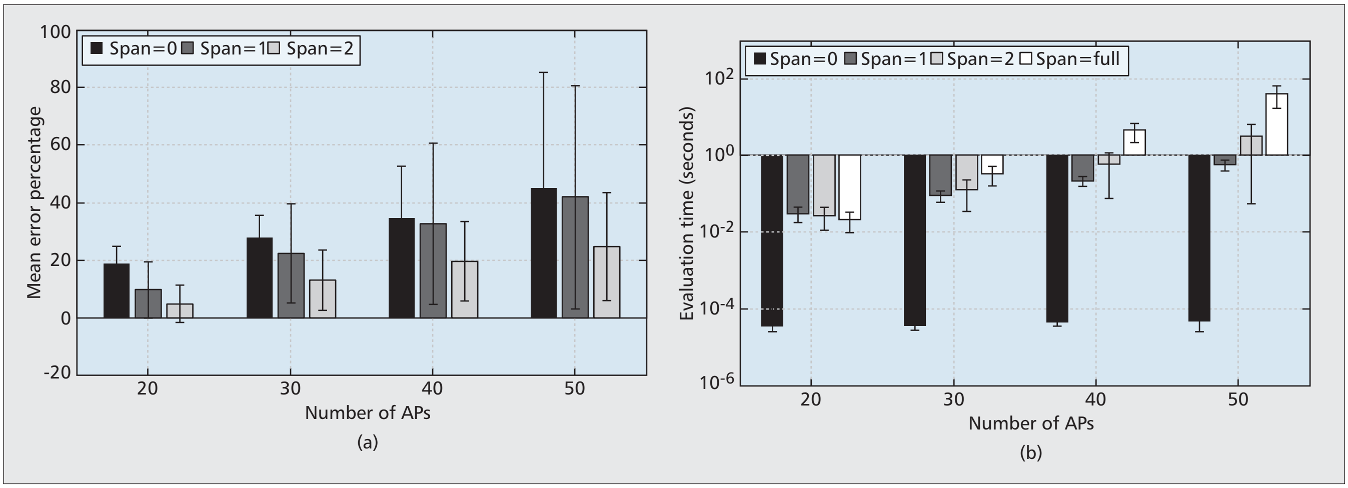


Figure 3. Performance of the approximation algorithm for different spans: a) error percentage compared to max span; b) evaluation times for different spans.

assignment scheme can result in a starved node if the neighborhood contention graph around the node is such that the node receives a much smaller share of the channel. We want to emphasize that although the MIS model would estimate a zero throughput for such a node, it is only applicable in scenarios where all nodes have saturation traffic over a long period of time. Since in reality some nodes may be intermittently idle, these starved nodes might get access to the channel during the idle-times of other nodes. Nonetheless, from a deployment perspective, the starved nodes identified by the model would be topologically vulnerable to performance problems and would offer very low throughput to connected clients during times of peak traffic, i.e. near-saturation load.

Simulation Results: Homogeneous Settings

We first benchmark the performance of three channel assignment strategies in homogeneous settings, i.e. all nodes follow the same algorithm:

- Random channel: each AP independently selects one of the three available channels respectively.
- Local selection: APs are deployed sequentially, and each AP selects the least congested channel from a local viewpoint.
- Centralized assignment: a single entity assigns the channel for all APs using a commonly used, greedy graph coloring heuristic [9].

In most real-world scenarios the number of available channels are far fewer than that required for completely conflict-free coloring. Thus this heuristic employs a multi-pass approach in which every pass involves identification of the most "saturated" node, i.e. the node with the largest number of already colored neighbors, and then assigns it a color that is least used among its neighbors. These three strategies are an extremely small subset of a vast trove of research as well as production algorithms for channel selection in wireless networks, including distributed schemes that specify optimal decisions for individual APs without the need for explicit cooperation [10]. However, these can serve as simplified representative strategies that occupy very different points in the space of possible channel assignment algorithms.

Figure 4a shows the mean normalized throughput at an AP for the three channel assignment schemes listed above. Each point shown in the plots is the average of 1000 simulations runs with AP locations chosen from a uniform random distribution within the simulation area of 1 sq. km. for each run. An interesting insight from this result is that a simple random channel selection performs reasonably well, especially in extremely dense settings since the gains from an optimal choice of channel is vastly reduced if all channels are almost equally crowded. However, in moderate densities (100–200 APs/sq.km), the centralized algorithms result in sizable gains of up to 30 percent. The gains from the centralization can be seen more prominently in terms of the starved node metric. Figure 4b shows the mean percentage of starved nodes (out of all nodes in the simulation) for varying densities. The performance of the random and local assignment schemes as per this metric generally follows the same trends as observed in Fig. 4a. However, the centralized assignment results in substantially fewer starved nodes at all densities.

Simulation Results: Mixed Settings

Next we consider deployments where different APs in range of each other use different channel assignment schemes. In reality, the number of different channel selection algorithms is bounded only by the number of different vendors (we observed more than 500 different vendors in the WiGLE.net dataset used in Fig. 1, and hence a myriad of scenarios with various permutations of AP locations and channel assignment schemes can arise). To make this analysis tractable for simulations, we compare the scenarios in which all APs under consideration either set channels based on a single centralized scheme or follow a different scheme independently. In practice, this assumption translates to the case of a single regional service assigning channels to all APs in

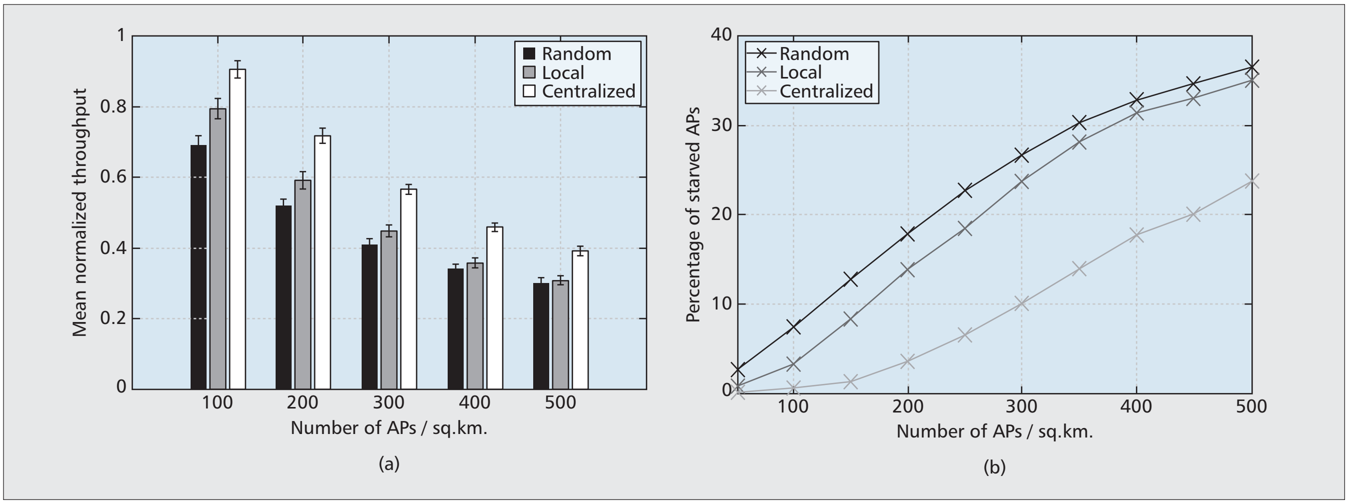


Figure 4. Comparison of channel assignment schemes in homogeneous settings: a) normalized throughput; b) starved nodes.

the region, except for a varying number of nonsubscribers. Alternatively, this is also applicable when the same fraction of APs implement a distributed version of a centralized algorithm by cooperating through a database service such as the TV White Space database [11].

The ratio of the number of APs following an independent scheme to those under the centralized scheme is varied from all APs belonging to one camp to all APs belonging to the other camp in step size of 5 percent of the total APs in consideration. For each ratio, 1000 simulation runs are performed where APs are deployed randomly and the independent group is chosen at random after the deployment. The centralized algorithm described above is used for the single centralized group, while three different assumptions are made for the independent group: random, local assignment, and same (all APs choose the same channel, for example channel 6 by default). Since the local assignment scheme involves scanning all channels locally for counting the number of neighboring APs on each channel, additional assumptions need to be made about the order in which the centralized and local assignments occur. For this, we assume that the APs in the independent group are turned on sequentially after the centralized group has fixed its channels. However, since we want the independent and centralized groups to reflect the behavior of low-cost residential APs and actively-managed hotspot APs, respectively, we assume that the local assignments, once made, remain fixed, while the centralized assignments are re-computed after the deployment of the local group.

Figure 5 shows both metrics described for the cases of random, local, and same channel assignments for the independent group. For each plot in these figures, the averages are computed over all the APs in the simulation, i.e. APs from the independent group and the centralized group together, and the shaded regions show the standard deviation around the mean values. The trends across all the cases are similar. There is a gradual increase in performance, in terms of both throughput and number of starved nodes as the ratio of APs acting independently is decreased. In other words, when deployment scenarios evolve from completely independent operation to a completely cooperative regime, throughput gains of the order of 40 percent and 15 percent are possible for the random and local assignments, respectively. The gains in terms of alleviating starved nodes are more pronounced, approximately 4x and 3x, respectively, for the same scenarios as above. Another interesting point to note here is that the performance of the centralized algorithm falls very gracefully in the presence of an increasing number of APs that are outside its control, as observed from the smooth nature of all curves. The extreme scenario of all independent nodes choosing the exact same channel is shown in Figs. 5e and 5f. As can be expected, the gains from all nodes using a centralized algorithm compared to individual operation are more here, approximately 2x in terms of mean throughput and 9x in terms of percentage of starved nodes.

The results above suggest that if the deployment trends shown in Fig. 1 continue, i.e. the percentage of more actively managed cooperating APs increases, the overall performance of APs will improve. However, when the same results are broken into the performance of the independent APs and the cooperating APs measured separately, a more nuanced view emerges. Figure 6a shows the breakup of the mean throughput between the two groups for a particular simulation: a mix of centralized and same channel APs with a density of 200 APs/sq. km. This shows that for a given density, as the percentage of independent APs decreases, the performance of the centralized APs also decreases, whereas that of the independent APs increases. This somewhat counter-intuitive result arises from the fact that when only a few APs make a smart choice about the channel in the presence of many "dumb" APs, they get more room to optimize the channel selection process. In other words, the worse performance of the independent APs in a setting with mostly independent APs is partially due to inefficient crowding of these APs onto certain channels, leaving more channels open to those APs that can sense and decide, whereas when only a few APs make sim-

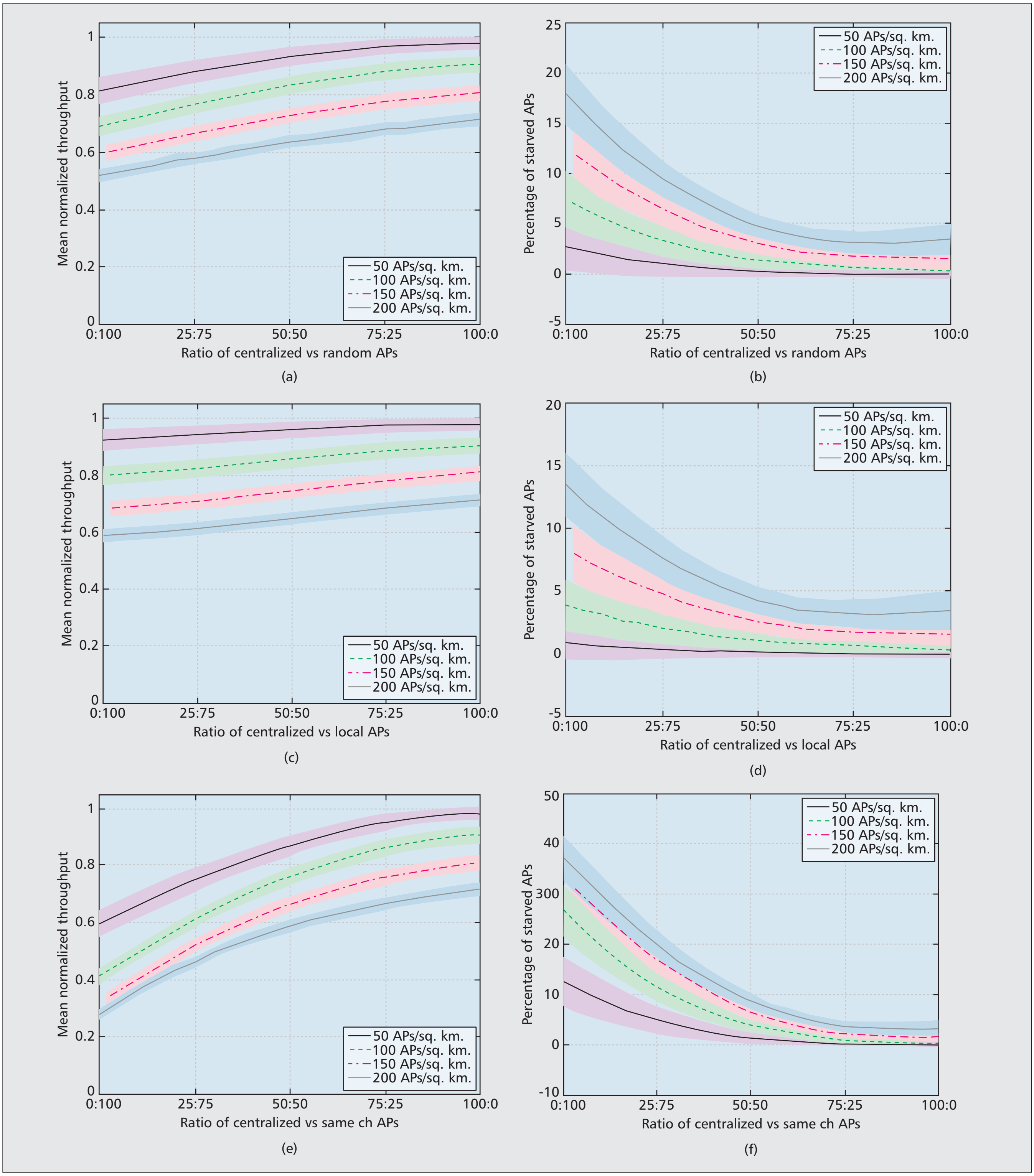
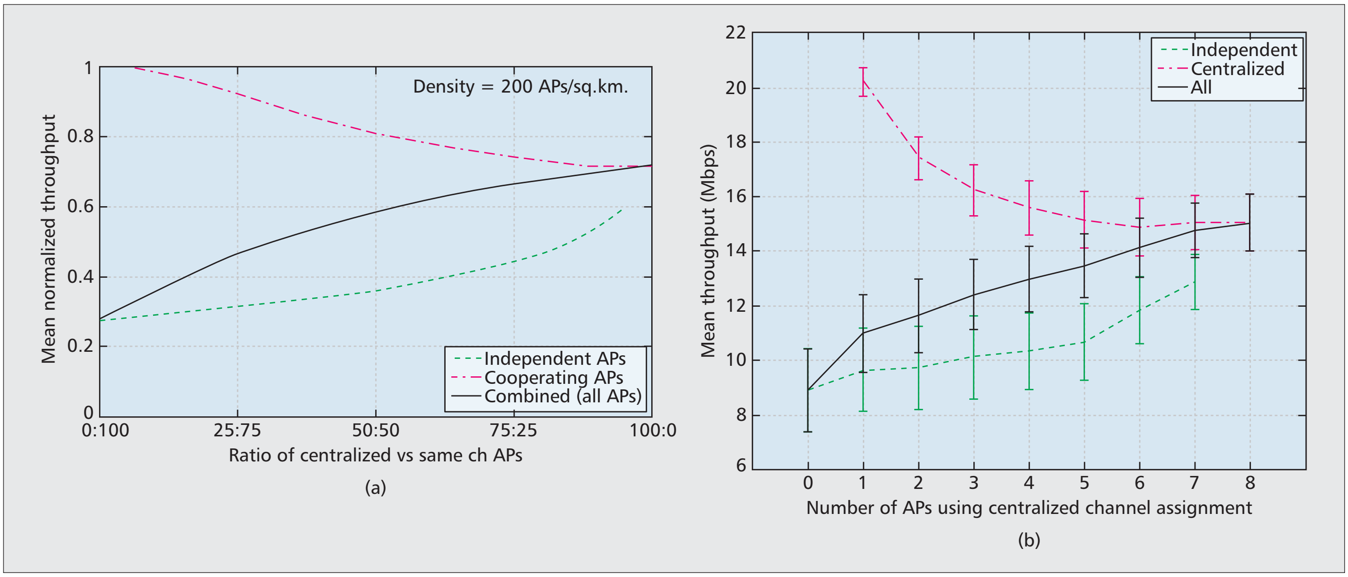


Figure 5. Performance under mixed deployment scenario. Figures a) and b): centralized vs. random; c) and d): centralized vs local; e) and f): centralized vs. same.

ple, static choices, the penalty of those choices is less severe since the centralized APs can sense and adjust their own channels accordingly.

Experimental Validation of Results

We performed a set of experiments with eight hardware nodes in order to verify the relation between the percentage of centralized nodes and their performance, observed in Fig. 6a. We use an eight-node attenuator system available as a part of the ORBIT lab facility [12]. This measurement system consists of eight Linux boxes, each of which has an Atheros 5212/5213 mini-PCI card and an Intel 6250 mini-PCIe 802.11/802.16 card. The nodes are enclosed in an RF enclosure that provides 80 dB of isolation,



(a)

(b)

Figure 6. Breakdown of the performance gain — increasing percentage of centralized APs leads to increasing performance of the independent APs: a) simulation results; b) experimental validation.

whereas all the input/output ports of the wireless cards are connected through a programmable attenuator. This setup provides a way to create arbitrary topologies (within the operating range of the attenuators) in a stable manner. Further details about this setup are available at [13].

For these experiments, we first randomly select 100 topologies out of the 11,117 connected topologies that are possible in a graph of exactly eight nodes [14]. The reason for this subsampling is that each experiment takes a considerable amount of time and the error margins obtained by averaging over 100 topologies seemed within acceptable bounds. For each topology, we run nine different experiments in which we vary the number of APs, choosing a constant fixed channel from 0 to 8 in increments of 1. Channels for the remaining APs in each experiment are assigned using the centralized-MIS scheme.

Figure 6b shows the performance of independent nodes, centralized nodes, and that of all nodes combined. As observed in the simulation results, this figure indicates that as the fraction of nodes that are under centralized control increases, the room for improvement in their performance decreases. Specifically, across 100 different topologies, the throughput obtained by a single "smart" AP in presence of seven other APs, all of which select the same channel, is on average 20 Mbps, whereas when all eight APs are under the same centralized channel assignment scheme, the average throughput of each AP is about 15 Mbps, indicating a ~25 percent drop in performance.

Implications and Discussions

In this paper we have highlighted the problem of co-existence between low-cost residential APs and actively-managed service provider APs in dense urban deployments. While several past works on heterogeneous radios have focussed on the interaction of different transmission technologies such as Wi-Fi, Bluetooth, and ZigBee, we study one specific difference that arises within Wi-Fi APs: the manner in which their channels are set. The vast difference in the cost and complexity of different APs poses a problem of unequal and unfair distribution of resources, especially in dense settings where performance is largely dependent on the number and type of other devices in the vicinity. We provide a detailed description of the simulation setup in order to encourage further studies of macro-level characteristics in dense wireless networks. Such simulations require abstraction of a number of finer aspects of the wireless medium, but enables the study of inter-linked mechanisms that can sometimes reveal counter-intuitive results. The simulation results presented in this paper lead to the following broader perspective and discussion points.

- Liew's MIS model for the channel share of a CSMA node, and the approximation of that model, shows that the performance of Wi-Fi APs can be greatly affected by the specific topology that happens to have formed due to other APs in its vicinity. The presence or absence of a single AP a few hops away can change the maximum independent sets of the graph and consequentially the channel share of an AP. Limiting the *span* around each node in the model can reduce the computation time but leads to larger approximation errors.
- In dense wireless network studies, starved nodes, i.e. nodes that are topologically vulnerable to performance problems, could be an important metric of interest. While under realistic traffic conditions these nodes might achieve nonzero throughput, the system design should aim at minimizing the number of such nodes since they will be prone to low channel share as and when traffic nears saturation.
- Increasing density of APs leads to substantial reduction in the average performance of channel selection schemes, and at 400-500 APs/sq.km. there are no benefits of a local chan-

nel selection strategy since random selection works equally well. Centralized channel assignment, on the other hand, helps at all densities, especially in terms of the starved node metrics.

- While the increase in the number of APs leads to lower performance for all APs, the residential APs would in fact be better off if new APs in their vicinity use a centralized infrastructure for channel planning, as compared to the new APs also being residential APs with fixed/local channel selection schemes.

An important extension of the problem as well as the simulation setup would be to consider multiple disjoint networks, each using the same or different centralized channel assignment algorithms. In general, it is not obvious to predict whether performance will still be close to the case of a single central agency, or will resemble that of local independent selections. Further work in this regard can also consider other radio resource allocation problems such as rate control, power control, and client-AP association optimization.

References

[1] S. Gundavelli *et al.*, Service Provider Wi-Fi Services over Residential Architectures, IETF Internet Draft, draft-gundavelli-v6ops-community-wifi-svcs-06.txt, Apr. 2013.

[2] WiGLE: Wireless Geographic Logging Engine, http://www.wigle.net/.

[3] Visualization of Wi-Fi Deployment over the Years for Different Cities in the US, http://www.winlab.rutgers.edu/baid/apViz.html.

[4] S.-C. Liew *et al.*, "Back-of-the-Envelope Computation of Throughput Distributions in CSMA Wireless Networks," *IEEE Trans. Mobile Computing*, vol. 9, no. 9, 2010, pp. 1319–31.

[5] L. Jiang and J. Walrand, "A Distributed CSMA Algorithm for Throughput and Utility Maximization in Wireless Networks," *IEEE/ACM Trans. Net.*, vol. 18, no. 3, June 2010, pp. 960–72.

[6] I. M. Bomze *et al.*, "The Maximum Clique Problem," H*andbook of Combinatorial Optimization*, Kluwer Academic Publishers, 1999, pp. 1–74.

[7] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE JSAC*, vol. 18, no. 3, 2000, pp. 535–47.

[8] A. Mishra, S. Banerjee, and W. Arbaugh, "Weighted Coloring based Channel Assignment for WLANs," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 9, no. 3, July 2005, pp. 19–31.

[9] J. Riihijarvi, M. Petrova, and P. Mahonen, "Frequency Allocation for WLANs Using Graph Colouring Techniques," *Proc. 2nd Annual Conf. Wireless On-demand Network Systems and Services (WONS)*, 2005, pp. 216–22.

[10] D. Leith and P. Clifford, "A Self-Managed Distributed Channel Selection Algorithm for WLANs," *4th Int'l. Symp. Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks,* Apr. 2006, pp. 1–9.

[11] "In the Matter of Unlicensed Operation in the TV Broadcast Bands: Third Memorandum Opinion and Order," Federal Communications Commission, Apr. 2012.

[12] D. Raychaudhuri *et al.*, "Overview of the ORBIT Radio Grid Testbed for Evaluation of Next-Generation Wireless Network Protocols," *Proc. IEEE WCNC*, 2005, vol. 3, pp. 1664–69.

[13] Sandbox 4 at ORBIT, http://bit.ly/1eeYi2w.

[14] Information System on Graph Classes and Their Inclusions, http://www.graphclasses.org/smallgraphs.html#nodes5.

Biographies

AKASH BAID (baid@winlab.rutgers.edu) is a network software engineer at Tarana Wireless. This work was part of his Ph.D. research at the Wireless Information Network Laboratory (WINLAB) at Rutgers University. He received his B.Tech degree in electronics and communications engineering from the Indian Institute of Technology Guwahati in 2008, and the M.S. and Ph.D. degrees in electrical and computer engineering from Rutgers University in 2010 and 2014, respectively.

DIPANKAR RAYCHAUDHURI [F] (ray@winlab.rutgers.edu) is a distinguished professor of electrical & computer engineering, and director of WINLAB (Wireless Information Network Lab) at Rutgers University. Dr. Raychaudhuri has previously held corporate R&D positions in Iospan Wireless, NEC USA C&C Research Laboratories, Broadband Communications Research, and Sarnoff Corp. He obtained his B.Tech (Hons) from the Indian Institute of Technology, Kharagpur in 1976, and the M.S. and Ph.D. degrees from SUNY, Stony Brook in 1978 and 1979, respectively.

While the increase in the number of APs leads to lower performance for all APs, the residential APs would in fact be better off if new APs in their vicinity use a centralized infrastructure for channel planning, as compared to the new APs also being residential APs with fixed/local channel selection schemes.

Per-Node Throughput Enhancement in Wi-Fi DenseNets

Kyungseop Shin, Ieryung Park, Junhee Hong, Dongsoo Har, and Dong-Ho Cho

Kyungseop Shin, Ieryung Park, Dongsoo Har, and Dong-Ho Cho are with Korea Advanced Institute of Science and Technology.

Junhee Hong is with Gachon University.

ABSTRACT

Wi-Fi networks are widely deployed for provision of Internet-centric data services. Since the inception of the Wi-Fi network in 1997 with its technical specification rooted in the IEEE 802.11 standard, much progress for higher data throughput has been made. Currently popular IEEE 802.11n Wi-Fi network in 2.4/5 GHz can deliver 600 Mb/s over a 40 MHz channel, which works well for most types of Internet-centric data services, and a later version of a Wi-Fi network based on IEEE 802.11ac is able to transmit at about 7 Gb/s. A simple configuration of a Wi-Fi network consisting of an AP and multiple stations for bidirectional data transmission enables low-cost implementation. High data rate provided at low cost as well as the abundance of Wi-Fi-capable mobile stations recently led to dense deployment of Wi-Fi networks, particularly in residential areas, business offices, and indoor/outdoor hotspots. However, dense deployment of Wi-Fi networks (e.g., Wi-Fi DenseNets) causes significantly increased overall interference, and as a result a significantly lowered achievable data rate. Thus, it is sensible to consider technologies that can resolve or mitigate deteriorated throughput of Wi-Fi DenseNets. In this article, technologies to deal with throughput enhancement of Wi-Fi DenseNets are addressed from three different perspectives: exploiting cellular technology for data transmission, elevating spectral efficiency, and controlling overall interference levels. Evaluation of interference control for Wi-Fi DenseNets is carried out in this article, and it is found that significant per-node throughput enhancement can be achieved.

INTRODUCTION

A typical Wi-Fi network consists of an access point (AP) and nodes (or stations, STAs) wirelessly connected to the AP. An AP and its associated nodes are called a basic service set (BSS). STAs such as smart phones, laptops, and smart pads can take on the role of AP to other STAs, establishing peer-to-peer (p2p) links between STAs in ad hoc mode [1]. This changeable configuration of a Wi-Fi network provides great flexibility in delivering different types of data services between APs and nodes or between nodes. Diverse applications requiring high data rate and abundant mobile STAs need more APs to attain concurrent data transmissions. With the typical service range of an AP covering tens of meters, Wi-Fi networks with respective APs are subject to dense deployments, particularly in airports, train stations, sports stadiums, apartments, and other hotspots. However, tens or hundreds of STAs within a small space of dense Wi-Fi networks, requesting diverse types of data services simultaneously, incur data collisions. Dense deployment of Wi-Fi networks naturally increases interference levels, so per-node throughput (or area throughput) of Wi-Fi DenseNets existing in such circumstances would be significantly decreased.

Recently, Qualcomm Inc. and Huawei Technologies Co. Ltd. announced the introduction of Long Term Evolution-Advanced (LTE-A) technology to the unlicensed frequency band (LTE-U) around 5 GHz [2]. The main goal of LTE-U is to offload explosively growing cellular traffic to the comparatively inactive unlicensed frequency band. Ironically, it is also conceived that data traffic of Wi-Fi DenseNets can be offloaded to cellular networks to get less dense operating conditions for Wi-Fi DenseNets [3].

Elevation of spectral efficiency is another option to mitigate the harsh operating conditions of Wi-Fi DenseNets. Current attempts to increase the spectral efficiency of 4G cellular networks [4] based on orthogonal frequency-division multiplexing (OFDM) can be adopted for Wi-Fi DenseNets. New technologies being considered for 5G cellular networks include non-orthogonal multiple access (NOMA) [5], spectrally efficient frequency-division multiplexing (SE-FDM) [6], and orthogonal frequency-division multiple access with variable tone spaces (OFDMA-VTS) [7]. Basic principles of these technologies are explained later in this article.

To deal with the issue of Wi-Fi DenseNets, the High Efficiency WLAN Study Group (HEW SG) was formed in May 2013, and as an extension of their activity, effort on standardization of IEEE 802.11ax was initiated in May 2014. The goal of the Task Group on IEEE 802.11ax is to improve per-node throughput of Wi-Fi DenseNets in the presence of interfering sources. They have considered indoor and outdoor environments such as public hotspots, apartments, and picocell streets, where Wi-Fi DenseNets are actually deployed. In order to increase area

0163-6804/15/$25.00 © 2015 IEEE

throughput of Wi-Fi DenseNets, they suggested dynamic sensitivity control (DSC) and transmit power control (TPC). The effect of DSC and TPC is evaluated, and the throughput gain from them is presented.

Technical evolution of Wi-Fi networks is described. Technologies to alleviate harsh operating conditions in Wi-Fi DenseNets and deployment scenarios devised by IEEE 802.11ax are also described. We present evaluation of DSC and TPC when applied to IEEE 802.11n DenseNets. We then briefly summarize this article.

Technical Review of Wi-Fi DenseNets

Figure 1 shows observed channel occupancy of IEEE 802.11n Wi-Fi networks in an indoor hotspot area of Suwon City, Korea. Figure 1a represents measured received signal strength indicator (RSSI) levels of Wi-Fi channels in 2.4 GHz band, and Fig. 1b illustrates measured RSSI level in 5 GHz band. Approximately 30 AP signals are observed in each frequency band, and each ID indicates an AP. For the majority of Wi-Fi channels, multiple APs coexist with similar RSSI levels, so associated AP and interfering APs cannot be judged by measured RSSI levels alone. It is seen in the figure that some Wi-Fi channels are overly crowded while other channels are not used at all. Uneven distribution of occupied Wi-Fi channels indicates that channels for APs are unfairly utilized. Wi-Fi DenseNets like this one can easily be found in other hotspots, and the number of detected APs even reaches 100 in some areas (e.g., the hotspot around Gangnam Subway Station in Seoul).

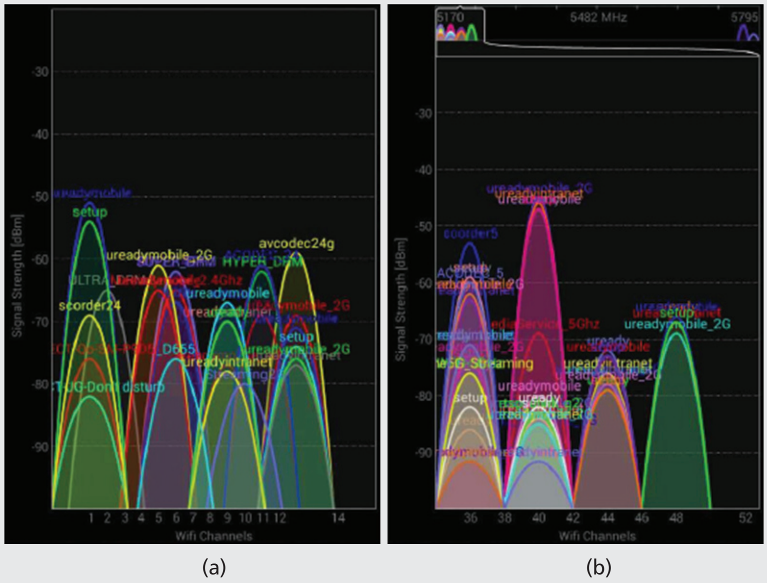


Figure 1. Dense deployment of 802.11nWi-Fi networks in a hotspot (snapshot obtained by Wifi Analyzer®): a) deployment of 2.4 GHz Wi-Fi DenseNets; b) deployment of 5 GHz Wi-Fi DenseNets.

Evolution of Wi-Fi Networks

The IEEE 802.11 standard family comprises multiple versions of Wi-Fi standards [1]. Starting from IEEE 802.11, which supports up to 2 Mb/s, the data rate of Wi-Fi networks has consistently increased. Currently popular IEEE 802.11n Wi-Fi networks for 2.4/5 GHz provide data rate up to 600 Mb/s with 4 spatial streams of multiple-input multiple-output (MIMO) data transmission. The IEEE 802.11n standard is backward compatible with the IEEE 802.11a/b/g standards. For high throughput, the medium access control/physical (MAC/PHY) layers of IEEE 802.11n support multiple spatial streams and other key features: aggregation of acknowledgment frames, aggregation of data frames, and reduction of inter-frame spacing. Techniques for improving spectral efficiency in 5 GHz band are mostly concerned with IEEE 802.11ac [8], and spatial multiplexing to support multi-user (MU) MIMO has been implemented in its MAC/PHY protocols. IEEE 802.11ac provides dynamic frequency selection, dynamic session transfer, beamforming, high order modulation up to 256 quadrature amplitude modulation (QAM), and simultaneous transmission of data frames in different access categories (e.g., different traffic classes, such as video and audio, belonging to MUs). The key feature of the 802.11ad [9] in 60 GHz band is directional MAC protocol essentially utilizing beamforming technology to overcome excessive path loss due to the high frequency of the carrier signal. Standardization activity for the IEEE 802.11 standard family has mainly focused on the enhancement of the link throughput in a single BSS, and has not provided a complete solution for overlapped BSSs (OBSSs). However, in reality, due to the low-cost deployment of Wi-Fi networks, APs are indiscreetly deployed, often creating OBSSs with aggravated spectral efficiency.

Types of Wi-Fi DenseNet Deployment

In the real world, Wi-Fi DenseNets are observed in several different types of environments. The IEEE 802.11ax Task Group designed DenseNets scenarios for indoor dense residences, indoor business offices, and indoor small BSS hotspots/outdoor large BSS hotspots [10]. The main purpose of such practical scenarios is to set common environments where techniques for Wi-Fi DenseNets are adopted for benchmark testing to prove their spectral efficiencies and per-node throughputs. Evaluation of tested techniques for DenseNets is valid only with OBSS. Figure 2 depicts three different types of DenseNet deployments. The residential scenario in Fig. 2a involves interference between APs placed in apartment units. A multi-story building with story height of 3 m is considered, and 20 units of 10 m × 10 m are located on a single floor. There is one randomly located AP per unit, and each unit has *N* uniformly (randomly) distributed STAs. Figure 2b shows an indoor enterprise Wi-Fi network on a single floor in an office building. There are 8 offices, and each office is 20 m × 20 m. In each office, 64 cubicles of 2 m × 2 m coexist, and each cubicle possesses 4 randomly distributed STAs. Also, for each office, four APs are installed. A scenario for indoor small BSS hotspot /outdoor large BSS hotspot is illustrated in Fig. 2c with frequency reuse factor 1. In the outdoor large BSSs hotspot scenario, cell coverage is shaped in a hexagon, approximating isotropic channel property, with an AP at the center, and the distance between adjacent APs set to 130 m. The BSSs in Fig. 2c, with their APs placed by an enterprise, can be called an enterprise service set. Small cells delimited by circles indicate small BSSs with standalone APs or p2p links. The standalone AP is an AP that is not managed by the enterprise of the enterprise service set. In this scenario, interference between APs in an enterprise service set, interference between STAs of p2p links, and interference between APs belonging to different enterprise service sets are considered. The indoor small BSS hotspot scenario is very similar to the outdoor scenario with some exceptions, such as much shorter distance (12 m) between adjacent APs and neglect of interference between APs belonging to different enterprise service sets.

Technologies for Wi-Fi DenseNets

Here, technologies recently considered to resolve technical challenges in Wi-Fi DenseNets are addressed from three different perspectives: exploitation of cellular technology for data transmission, elevating spectral efficiency, and controlling overall interference level. Simultaneous transmission of LTE-U and Wi-Fi networks will mitigate traffic overload problems of 5 GHz frequency band where Wi-Fi networks are only utilized, and non-orthogonal multiple access schemes might be able to enhance spectral efficiency. DSC and TPC can be used effectively to reduce the overall interference level, which is converted to increased per-node throughput.

LTE-U — The incremental capacity of cellular networks cannot keep pace with the growth of cellular traffic. This situation motivates cellular operators to migrate some portion of cellular traffic to another frequency band. In order to take over some traffic in the cellular network, IEEE 802.11u [1] suggested the Wi-Fi passpoint, which enables handoff from a cellular network to a Wi-Fi network. However, seamless handoff cannot be achieved by a Wi-Fi passpoint alone. Therefore, demand for more systematic offloading has increased. Lately, Qualcomm Inc. and Huawei Technologies Co. Ltd reported LTE-U based on

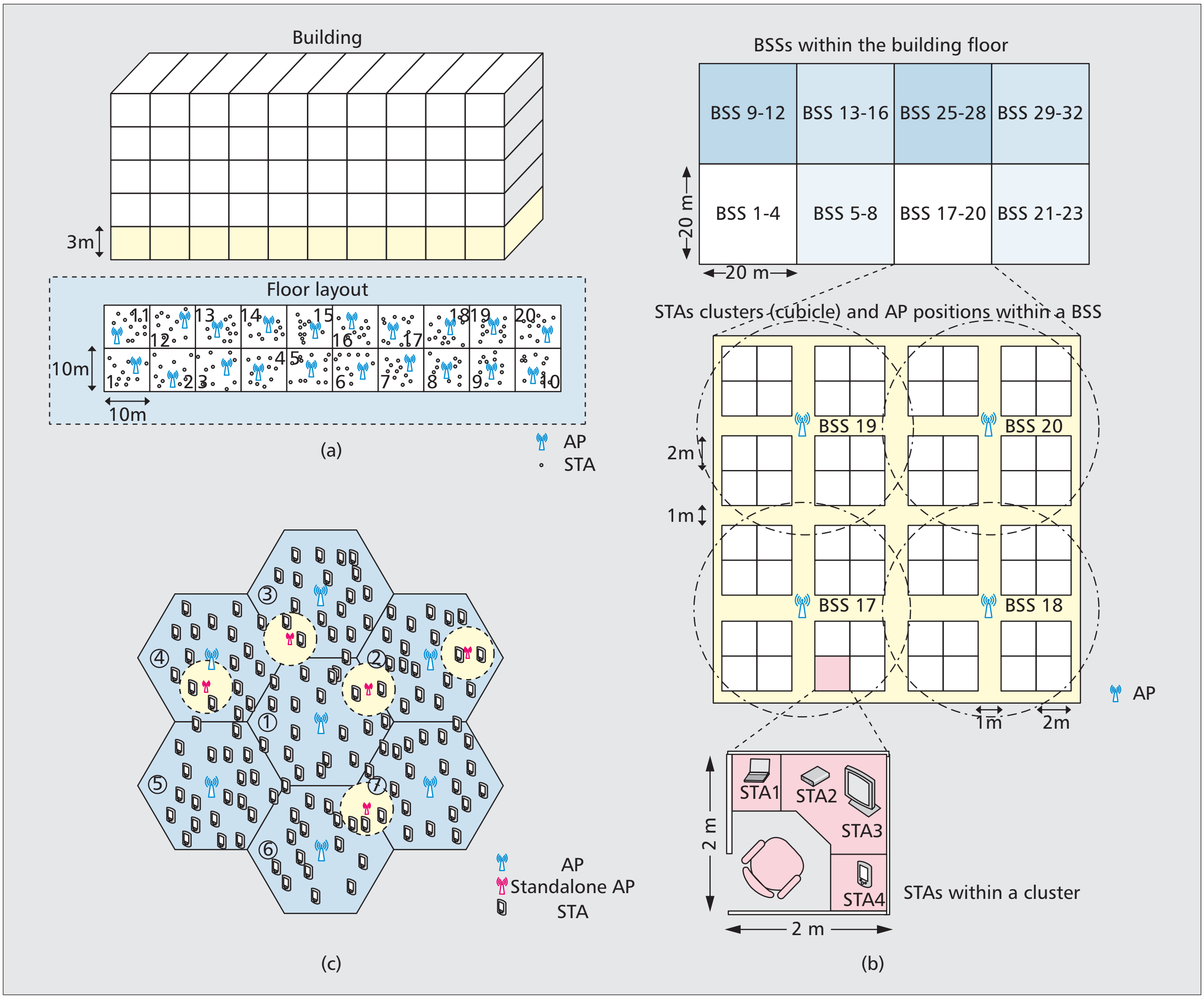


Figure 2. Deployment scenarios for Wi-Fi DenseNets: a) residential scenario; b) enterprise scenario; c) indoor/outdoor BSS hotspot scenario.

LTE-A. The main target of LTE-U is to alleviate large traffic of cellular networks by making use of unlicensed frequency band in 5 GHz. In Fig. 3, the aggregation of unlicensed carriers with licensed ones is illustrated. Both downlink and uplink are operated by LTE-A technology. For downlink transmission, carrier aggregation is performed to provide a higher data rate while a licensed carrier alone might be used for uplink transmission. Downlink transmission of data traffic is best effort type transmission, and in uplink, data transmission in unlicensed frequency band is allowed only when there is no significant interference detected or incurred during data transmission. Critical control frames for LTE-U are transmitted using licensed frequency band for both downlink and uplink. Unlicensed carriers are considered secondary carriers, and their operations are controlled by a cellular network. Joint scheduling between LTE-A and LTE-U is performed by the LTE-A cellular network. Currently, LTE-U and Wi-Fi networks are often discussed together for small cells and p2p communications.Cellular standardization activities to interwork with Wi-Fi networks have focused on handling traffic overflow of cellular network. In Third Generation Partnership Project (3GPP) standards, the interoperability (Releases 6 and 7), seamless handover and service reliability (Release 8 and 9), integrated access (release 10), and systematic interoperability of multiple radio access technologies (Release 13) between cellular and Wi-Fi networks are stated. The series of standardization efforts for interoperability between cellular and Wi-Fi networks enables smooth migration of data traffic from a cellular network to a Wi-Fi network, and vice versa. In the future, data traffic of Wi-Fi DenseNets can be handled in LTE-U to control the overall interference level of Wi-Fi DenseNets. Congestion of data traffic in unlicensed band for Wi-Fi devices might be resolved by simultaneous use of Wi-Fi and LTE-U technologies.

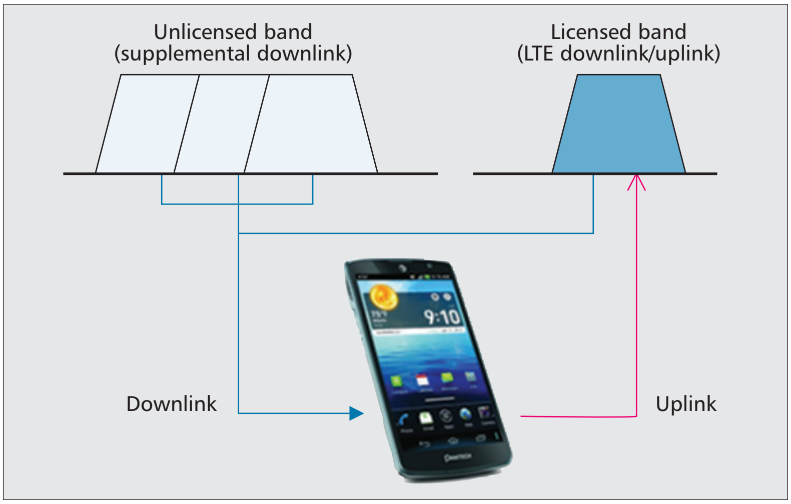


Figure 3. Downlink/uplink of LTE-U network exploiting unlicensed frequency band of Wi-Fi network.

Other Multiple Access Schemes: NOMA, SE-FDM, and OFDMA-VTS — For 5G cellular networks, non-orthogonal multiple access schemes to surpass the OFDM scheme in terms of spectral efficiency are under consideration. Most of these non-orthogonal multiple access schemes require higher system complexity and implementation cost than the OFDM scheme. However, it is expected that advanced signal processing and system-on-a-chip technology will help overcome these demerits. As non-orthogonal multiple access schemes, NOMA, SE-FDM, and OFDMA-VTS are explained.

The NOMA scheme in Fig. 4b exploits superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver. In downlink, if a station STA1 is located closer to an AP than other station STA2, the AP allocates more power to STA2 with power level P2, compared to STA1 with smaller power level P1. The signals transmitted from the AP are superposed, as seen in Fig. 4b, in the power domain over the same carrier. With more transmitting power for STA2, STA2 can decode its signal directly, treating the signal for STA1 as interference. On the other hand, STA1 decodes its signal in an iterative (successive) process. STA1 first decodes the signal for STA2, and reconstructs the signal for STA2 from decoded bits. The reconstructed signal is subtracted (cancelled) from the superposed signal, and the residual signal is decoded for STA1.This successive cancellation of interference for an STA can be extended for MUs. Provided that the NOMA scheme is adopted for OFDM, multiplexing gain over the power domain can bring improved spectral efficiency for OFDM systems.

The SE-FDM scheme in Fig. 4c uses tone spacing smaller than the tone spacing of the OFDM scheme (i.e. $\Delta f_S < \Delta f_O$). Decreased tone spacing with fixed total bandwidth indicates correspondingly increased spectral efficiency. The increased spectral efficiency of the SE-FDM is obtained at the cost of intentionally violated subcarrier orthogonality. The violated subcarrier orthogonality inevitably introduces interchannel interference (ICI), so the impact of the ICI must be reduced for successful decoding. To this end, high-complexity data decoding methods such as maximum likelihood detection and Gram-Schmidt orthogonalization are required.

The OFDMA-VTS scheme is based on variable tone spacing, as shown in Fig. 4d, so $\Delta f_1 \neq \Delta f_2$. Since the carrier frequency and mobility of each user are different, user-specific delay and Doppler frequency shift will be experienced by user subchannels. This leads to each user having unequal intersymbol interference (ISI) and ICI. Therefore, fixed tone spacing like OFDMA is inefficient for these user-specific subchannels, and results in cell capacity degradation. In OFDMA-VTS, the tone spacing for each subchannel is adjusted according to minimization of the composite interference (e.g., ISI plus ICI). Optimized tone spacing can provide increased spectral efficiency.

IEEE 802.11ax — Standardization activity of the IEEE 802.11ax Task Group, concerned with densely deployed WLANs, began in May 2014. Approaches appeared in the submissions of the HEW SG to enhance node throughput in DenseNets can be described as follows.

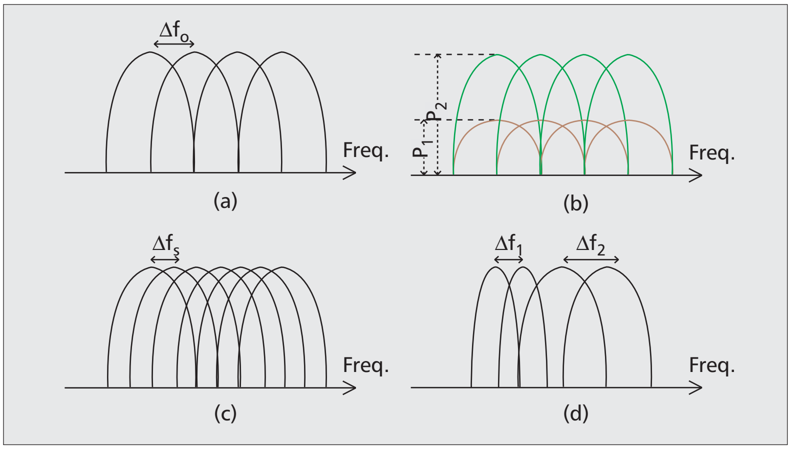


Figure 4. OFDM scheme and non-orthogonal multiple access schemes: a) OFDM; b) NOMA; c) SE-FDM; d) OFDMA-VTS.

"Possible Approaches for HEW" [11] deals with congestion, interference, and frame conflicts as three major problems in Wi-Fi DenseNets. The airtime of Wi-Fi DenseNets consists of frames, inter-frame space, and contention windows. Reduction of the control frames' overhead as well as aggragation of the data frames decreases the airtime of control frames. Congestion caused by excessive nodes taking up most of the airtime might be resolved by reducing the size of a control frame and increasing the relative size of a data frame. It is also mentioned that better quality of experience (QoE) can be obtained by limiting the number of associated STAs. Interference caused by dense deployment of Wi-Fi networks can be reduced by restricting the access of STAs located in the cell edge, which have comparatively low signal-to-interference-plus-noise ratio (SINR) and thus worse node throughput performance. Frame conflicts typically occur in hidden STAs and can be solved by modification of the channel access scheme and OBSS management.

"Enhancement on Resource Utilization in OBSS Environment" [12] suggests relaxed network allocation vector (NAV) protection to improve spatial reuse efficiency or per-node throughput. When an STA receives a request-to-send (RTS) or clear-to-send (CTS) signal, it sets NAV, which is a counter for how long the channel will be busy, and does not transmit any data. Thus, the STA does not attempt to transmit even if no CTS signal responding to the RTS is overheard. This would reduce spatial reuse efficiency. If the NAV is updated only when the CTS signal is overheard, STAs having data to transmit to other STAs can deliver data. From this viewpoint, relaxed NAV protection can increase spatial reuse efficiency.

Enhancing Per-Node Throughput in Wi-Fi DenseNets

In this section, enhancement of per-node throughput by the DSC and TPC is demonstrated, following the HEW SG scenarios. The main objective is to compare per-node throughputs of Wi-Fi DenseNets with and without interference control.

Per-Node Throughput Evaluation Condition

Clear channel assessment (CCA) is a channel sensing mechanism based on carrier sense and energy detection. The CCA sensitivity indicates a predefined threshold to judge the channel as busy or idle. The DSC dynamically changes the CCA sensitivity. Depending on the CCA sensitivity, the STA accesses the channel aggressively or passively. Benefits from the DSC can be explained in a couple of ways. When the interference level measured at an STA is close to the beacon signal level of an associated AP, the STA holds off data transmission to reduce overall interference. If there is a much lower interference level, the STA proceeds to transmit data to increase per-node throughput. TPC controls the transmit power at the transmitter to provide a predetermined power level at the receiver. Without TPC, the transmit power at the transmitter is fixed. When an STA is located close to the AP, the transmit power of the AP for the STA can be made small by leveraging small path loss of the link. This allows more STAs to establish p2p links within the BSS, and also provides a higher chance for APs and STAs in adjacent cells to reuse the same channel [13].

For the DSC, CCA sensitivity is set to (R-M) dBm, where R stands for the RSSI level of the beacon signal of the associated AP, and M represents a margin. M is set to 20 dB, which is a typical value for the margin. It is assumed that the channel characteristics over the time duration of the beacon signal transmission and data signal transmission by an AP are stationary. The TPC of the AP is adjusted to make the RSSI level of the receiving STA become –50 dBm, which is often considered a target RSSI level based on TPC [13]. For evaluation according to the scenarios in Fig. 2, downlink of IEEE 802.11n Wi-Fi networks is considered. The Tx power of AP is set to 18 dBm in a residential scenario, 21 dBm in an enterprise scenario, and 15 dBm in an indoor small BSS hotspot scenario [10]. The respective number of APs is 20 for residential, 32 for enterprise, and 7 for indoor small BSS hotspot. Each AP is equipped with four transmitting antennas, and each STA has two transceivers. Noise level is set to –101 dBm (= –174 dBm/Hz [thermal noise] × 20 MHz [channel bandwidth]), and SINR is evaluated for each STA.

Saturated payload for a given number of STAs is assumed, so the AP or transmitting STA of a p2p pair or standalone AP always has data to transmit. Also, collision-free scheduling, load balancing among STAs, and control overheads are considered. Other simulation parameters [10] include data frame size = 1472 bytes, aggregation level of data frames = 2, preamble duration = 40 μs, acknowledgement packet duration = 68 μs, RTS duration = 52 μs, CTS duration = 44 μs, SIFS duration = 16 μs, and expected waiting time for the channel acquisition by enhanced distributed channel access protocol = 100.5 μs. It is also noted that STAs do not act as APs for other STAs when they have connections to other APs.

Maximum achievable node throughput is obtained by Shannon's channel capacity formula for 4 × 2 MIMO systems. To get channel capacity by the formula, SINR is required. The interference level of SINR for an STA is evaluated over individual interfering APs, depending on their locations. Following the residential scenario, a single channel is used for all the APs. Categorization of interfering APs according to the number of penetrations of walls and locational proximity of each unit can be considered to simplify the computational process. If an STA in an apartment unit with ID = 3 in Fig. 2a, APs placed in adjacent units with ID = 2, 4, 13 and the other APs in remaining units are treated differently. For the APs in units with ID = 2, 4, 13, each of which shares a wall with the unit under consideration, combinations of STAs, one from each unit, are exhaustively taken into account. Let us assume that a composite interference level is obtained as $I_{adj} + I_{rem}$, where $I_{adj} = \Sigma I_a$, I_a = interference from an adjacent AP with index a, and $I_{rem} = \Sigma I_{na}$, I_{na} = interference from an AP with index na that does not belong to the set of adjacent APs. For every combination of nodes, the accumulated interference level is computed as I_{adj}. With the remaining units, the average interfering power of respective AP over STAs in each unit is computed and then added together over the remaining units to get I_{rem}. It is noteworthy that I_{rem} is used for all the STAs in the unit under consideration, whereas I_{adj} is evaluated for each different combination of STAs in adjacent units. With this composite interference level, the SINR and channel capacity of an STA are computed. Changing combinations of STAs in adjacent units, composite interference level and channel capacity are re-evaluated. Mean channel capacity obtained this way is assigned as node throughput. For other units such as that with ID = 11, only two adjacent units with ID = 1, 12 need to be considered for I_{adj}. From this process for STAs in units, per-node throughput is obtained. For the enterprise scenario, different channels are assigned for eight offices, so such categorization of interfering APs is not necessary. As depicted in Fig. 2b, four channels marked by distinct colors of offices are assigned, avoiding co-channel interference. Therefore, only three interfering APs and their associated STAs in the same office are considered as interference sources, neglecting co-channel interference because of distance. In the indoor BSSs hotspot scenario, all six interfering APs and their STAs are taken into consideration.

The IEEE 802.11n channel model [15] is adopted as the path loss model. Wall penetration loss is set to 12dB. Carrier frequency is 2.4 GHz with 20 MHz channel bandwidth. Evaluation of per-node throughput is performed by Matlab. Per-node throughput is acquired by averaging individual node throughputs of all the receiving nodes in a run, and the average per-node throughput in Fig. 5 with a 95 percent confidence interval is obtained over runs. The locations of APs in the residential scenario and STAs in all the scenarios are randomly updated in each run.

Results from DSC and TPC

Figure 5a shows the variation of per-node throughput according to the number of nodes without p2p pairs in an apartment unit(left side) and also according to the number of p2p pairs (right-hand side figure). Per-node throughput evaluation for p2p pairs is performed when the total number of nodes is 10 per apartment unit. Nodes for p2p pairs for p2p links are selected out of 10 nodes in each unit, and the remaining nodes are associated with an AP. It is noted that average per-node throughput is calculated for all the receiving nodes, including the receiving nodes of p2p pairs. With the DSC and TPC jointly used for downlink transmission, significant improvement of per-node throughput is obtained. The DSC enables an AP to get more transmission opportunities, since the nodes in the cell sense the channel as idle more frequently, and in addition to the presence of intervening walls that significantly reduce interference from other APs, the TPC is efficient in curtailing overall interference. Per-node throughput gain over the IEEE 802.11n networks without the DSC and the TPC is ranging from about 50 to approximately 100 percent, depending on the number of nodes.

Figure 5b demonstrates the variation of per-node throughput performance with increased number of nodes and p2p pairs per four cubicles, as depicted in Fig. 2b. The p2p pairs are selected out of 16 nodes. Since there is no wall between APs, the accumulated interference level due to adjacent APs is relatively large in comparison with the residential scenario, and as a consequence, gain of the per-node throughput by DSC and TPC is decreased in the enterprise scenario. This can be commonly observed in Figs. 5a and 5b with or without p2p pairs. Nevertheless, per-node throughput gain with or without p2p pairs is about 40 percent.

Figure 5c shows per-node throughput for indoor small BSSs hotspot scenario where seven BSSs coexist. Gain due to DSC and TPC is about 20 percent with a small number of nodes and increases up to 200 percent as the number of nodes increases. Compared to Fig. 5a, the effect of DSC and TPC in terms of gain is more profound with increased number of nodes or p2p pairs. This is partially due to the random locations of APs in a residential scenario, which can accentuate the adverse impact of some interfering APs close to the node under consideration. Less significant overlapping of cell coverage in the indoor small BSS hotspot scenario, compared to the significant overlapping of cell coverage in the enterprise scenario in Fig. 5b, seems to be beneficial to obtain the gain of DSC and TPC. The graph on the right side of Fig. 5c represents evaluation results according to the number of standalone APs when 30 nodes exist in each BSS. Each standalone AP is associated with two STAs. With standalone APs, gain over the IEEE 802.11n Wi-Fi network without DSC and TPC is greater than 100 percent. As a whole, gain from DSC and TPC for Wi-Fi DenseNets is significant for various types of deployments.

Conclusion

This article has focused on technologies suitable for Wi-Fi DenseNets. LTE-U has been considered as a candidate technology to offload data traffic of Wi-Fi DenseNets to cellular networks. Several non-orthogonal multiple access schemes

Clear channel assessment (CCA) is a channel sensing mechanism based on carrier sense and energy detection. The CCA sensitivity indicates a predefined threshold to judge the channel as busy or idle. The DSC dynamically changes the CCA sensitivity. Depending on the CCA sensitivity, the STA accesses the channel aggressively or passively.

for 5G cellular networks were addressed as potential approaches to increase spectral efficiency of Wi-Fi DenseNets. In general, enhancement of spectral efficiency is obtained at the cost of increased system complexity. Three different deployment scenarios suggested by the HEW SG have been applied to find out the benefits of interference-level control. An interference control scheme in the form of DSC and TPC, suggested by the HEW SG, has been tested with three different deployment scenarios. It has been found that control of CCA sensitivity and transmit power to reduce the overall interference level provides significant per-node throughput improvement to Wi-Fi DenseNets, regardless of deployment scenarios.

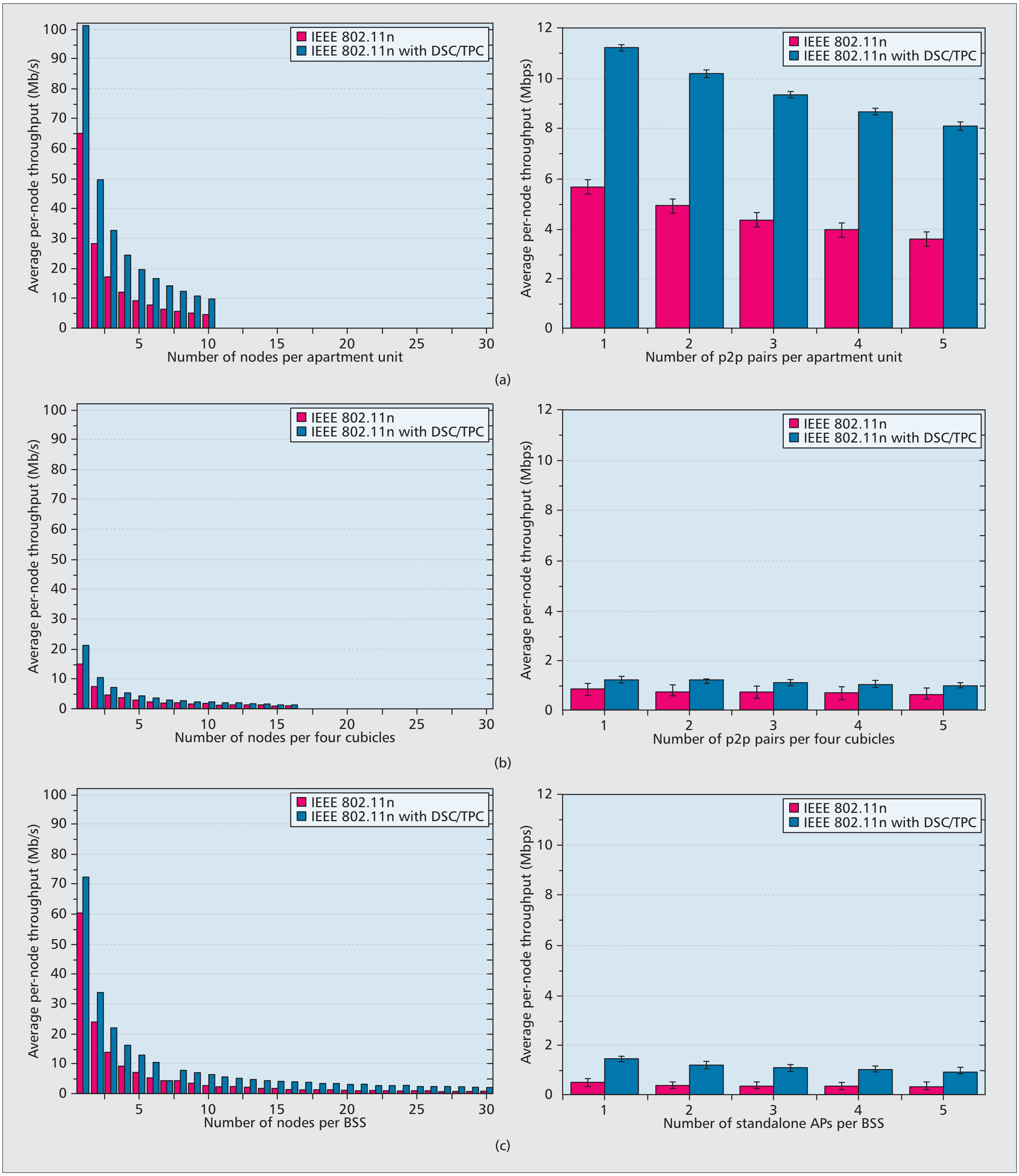


Figure 5. Evaluation results for Wi-Fi DenseNets according to the scenarios in Fig. 2: a) residential scenario; b) enterprise scenario; c) indoor BSSs hotspot scenario.

Acknowledgment

This research was supported by ICT R&D program of MSIP/IITP [11-911-04-001, Development of Adaptive Beam Multiple Access Technology without Interference based on Antenna Node Grouping]. Also, this research was supported by Korea Ministry of Land, Infrastructure and Transport (MOLIT) as Railroad Specialized Graduate School.

References

[1] IEEE 802.11-2012, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," Mar. 2012.

[2] "Extending the Benefits of LTE Advanced to Unlicensed Spectrum," QUALCOMM, Nov. 2013, http://www.qualcomm.com/media/documents/extending-benefits-lte-advanced-unlicensed-spectrum

[3] I. Hwang *et al.*, "A Holistic View on Hyper-Dense Heterogeneous and Small Cell Networks," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 20–27.

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

[5] NTT DOCOMO, "Requirements, Candidate Solutions & Technology Roadmap for LTE Rel-12 Onward," 3GPP RWS-120010, June 2012.

[6] I. Kanaras *et al.*, "Spectrally Efficient FDM Signals: Bandwidth Gain at the Expense of Receiver Complexity," *Proc. IEEE ICC*, June 2009, pp. 2701–3706.

[7] S. Lim *et al.*, "Optimal Tone Space Selection Scheme for OFDMA-VTS in Carrier Aggregation," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, Nov. 2013, pp. 5679–91.

[8] IEEE 802.11ac, http://www.ieee802.org/11/Reports/tgac_update.htm.

[9] IEEE 802.11ad, http://www.ieee802.org/11/Reports/tgad_update.htm.

[10] IEEE 802.11-14/0621r3, "Tgax Simulation Scenarios," Sept. 2014.

[11] IEEE 802.11-13/0758r0, "Possible Approaches for HEW," July 2013.

[12] IEEE 802.11-13/1056r1, "Enhancement on Resource Utilization in OBSS Environment," Sept. 2013.

[13] IEEE 802.11-13/1290r1, "Dynamic Sensitivity Control for HEW SG," Apr. 2014.

[14] IEEE 802.11-03/940r4, "TGn Channel Models," May 2004.

Biographies

KYUNGSEOP SHIN received his B.Sc. and M.Sc. degree from KAIST, Daejeon, Korea in 2009 and 2011. Now he is in the Ph.D. program under the supervision of Prof. Dong-Ho Cho. His current research interests are bioinformatics, and medium access protocol for ad hoc wireless communications and wireless local area networks.

IERYUNG PARK received his B.Sc. from Kyunghee University in 2007 and his M.Sc. degree from Gwangju Institute of Science and Technology in 2009. He has been involved with a variety of industrial projects on communication system design at Samsung and CoreLogic. Now he is in the Ph.D. program under the supervision of Prof. Dongsoo Har. His current research interests are machine-to-machine communications and wireless power transfer.

JUNHEE HONG received his B.Sc. degree in electrical engineering from Seoul National University, Korea, in 1989 and his Ph.D. degree in electrical engineering from Seoul National University in 1995. He has been a professor of electrical power engineering at Gachon University, Kyunggi-do, Korea, since 1995. His research interests include smart grid, super grid, renewable powered desalination, and their communication infrastructure.

DONGSOO HAR received his B.Sc. and M.Sc. degrees in electronics engineering from Seoul National University in 1986 and 1988, respectively. He received his Ph.D. degree in 1997 from Polytechnic University, Brooklyn, New York. He was awarded Best Paper Award (Jack Neubauer Award) from *IEEE Transactions on Vehicular Technology* in 2000. His microcell model (Har-Xia-Bertoni model) has been widely quoted for practical deployment of cellular networks. Since 2013 he has been a faculty member of KAIST.

DONG-HO CHO [M'85, SM'00] received a Ph.D. degree in electrical engineering from KAIST in 1985. Since 1998, he has been a professor in the Department of Electrical Engineering of KAIST. He has been a director of the KAIST Online Electric Vehicle Project since 2009 and has served as head of the Cho Chun Shik Graduate School for Green Transportation since 2010. His research interests include mobile communication, online electric vehicle systems based on wireless power transfer, and bioinformatics.

It has been found that control of CCA sensitivity and transmit power to reduce the overall interference level provides significant per-node throughput improvement to Wi-Fi DenseNets, regardless of deployment scenarios.

On the Efficient Utilization of Radio Resources in Extremely Dense Wireless Networks

Arash Asadi, Vincenzo Sciancalepore, and Vincenzo Mancuso

Arash Asadi is with IMDEA Networks Institute and the University Carlos III of Madrid

Vincenzo Sciancalepore is with IMDEA Networks Institute, the University Carlos III of Madrid, and the University Politecnico di Milano.

Vincenzo Mancuso is with IMDEA Networks Institute and the University Carlos III of Madrid.

ABSTRACT

The emergence of popular wireless technologies such as LTE and WiFi, and the exponential growth in the usage of these technologies, has led to extremely dense wireless networks. There are many proposals for coping with such densification. In particular, we evaluate the compound effect of inter-cell interference schemes and spectrum efficient intra-cell relay techniques, which have been individually proposed recently as separate solutions. We provide a jointly coordinated intra-cell and inter-cell resource allocation mechanism that opportunistically exploits network density as a resource. We show that intra-cell opportunistic relay, based on WiFi communications, reduces the complexity of Inter-Cell Interference Coordination (ICIC) and boosts the efficiency of ICIC in LTE. The superiority of the proposed solution to the legacy cellular network operation is proven via simulations.

INTRODUCTION

Wireless networks are densifying rapidly due to the widespread emergence of wireless technologies in consumer devices (e.g. mobiles, laptops, and house appliances), the rapid growth of data-driven applications in the market (e.g. online games and social networks), and high adoption of these applications by users [1]. While this densification indicates the popularity of wireless technologies, it also introduces pressing issues such as interference, and spectral and energy efficiency. These issues also exist in the legacy wireless networks, but they are mostly circumvented by weakening their impact on the system performance. Interestingly, all the aforementioned issues are related to the use of wireless spectrum and require an efficient approach to the utilization of radio resources. This goal can be achieved by using suitable mechanisms for interference mitigation/control and efficient use of the spectrum. In this article we give a concrete example of such integration, with particular focus on interference mitigation and opportunistic channel utilization.

The existing solutions for the aforementioned issues are abundant. The interference issue is usually addressed by using interference mitigation techniques such as coordinated beamforming, power allocation, and Cooperative MultiPoint (CoMP). For instance, fast distributed beamforming in multi-cell environments has been proposed in [2], in which scheduling is performed in two steps:

- Each base station chooses the proper beamforming pattern in order to minimize the inter-cell interference.
- A particular set of users is scheduled in each cell.

Additionally, valid heuristics have been designed to properly allocate the resource blocks when adjacent cells interfere with each other [3, 4]. These approaches avoid the interference of the two most interfering base stations by allocating the cell-edge users (where the interference is proved to be significant) on different resource blocks. Graph theory is another tool for modeling network interference in CoMP mechanisms. The authors of [5] propose a graph coloring technique for interference coordination that is based on two interference graphs. The first graph (*outer graph*) uses global per-user interference information; the second graph (*inner graph*) takes advantage of local information obtained from the base station, and global constraints derived from the global graph. Recently, 3GPP has standardized a new technique, called ABSF (Almost Blank Sub-Frame[1]) that assigns resources in such a way that a subframe may be blanked for some base stations in order to prevent their activity in high interference scenarios. Some of the inter-cell interference coordination mechanisms leverage the ABSF in order to improve the spectral efficiency of the network by reducing the global interference [6].

Although interference mitigation implicitly improves spectral efficiency, some researchers explicitly aim to improve the spectral efficiency of dense networks by exploiting new resource allocation methods that leverage the high user density. For example, the authors of [7, 8] propose integrating opportunistic scheduling and a popular branch of cooperative communications, namely Device-to-Device (D2D) communication, to enhance the spectral efficiency of the networks by leveraging dynamic clusters of users.

[1] 3GPP, "Evolved Universal Terrestrial Radio Access Network (EUTRAN); X2 application protocol (X2AP), 3rd ed" 3GPP, technical Specification Group Radio Access Network, Tech. Rep.

0163-6804/15/$25.00 © 2015 IEEE

Interestingly, this approach also improves the energy efficiency in dense networks by increasing transmission rate and reducing the power spent for keeping the wireless interface in active mode. Many of the above mentioned solutions require small modifications in the operations or infrastructure of the network. However, the rigidness of the current infrastructure does not allow these modifications without going through the lengthy standardization process. Software Defined Networking (SDN) is proposed to turn the current rigid network structure into a flexible network. Although SDN paves the way toward implementing a variety of enhancing techniques for dense network, one should carefully choose these techniques as some of them might act on common network features and parameters. In general, a thorough solution for dense wireless network provides the means for:

- Controlling inter-cell interference caused by dense base station implementations.
- Improving spectral and energy efficiency inside the cell.
- Creating a flexible architecture that accommodates the first two while allowing interoperability among different platforms (e.g. LTE and WiFi).

In our work we select ABSF, D2D-enabled opportunistic scheduling, clustering, and SDN techniques, which satisfy all the requirements of a comprehensive network solution. To improve efficiency of the resource utilization in dense wireless networks, we present the solution developed in the frame of the CROWD project,[2] which targets frequency reuse to maximize the use of licensed spectrum in the network. We show how the combination of intra-cell resource optimization (achieved via D2D and clustering techniques), and inter-cell interference control (achieved via ABSF) is feasible and suitable for dense scenarios.

Network Densification: Issues and Solutions

Dense wireless networks inherit the same issues of the legacy cellular system. It is the magnification of these issues that demands solutions specific for dense networks. In what follows, an overview of the issues and feasible solution for dense network is provided.

Issues

Interference — The cellular technology manufacturers counterbalanced the intensive demand with implementation of micro cells and femto cells to increase the frequency reuse and hence spectral efficiency, as shown in Fig. 1. However, this approach also exposes the system to more interference. In general, cellular communication is exposed to two major sources of interference, namely, intra-cell interference and intercell interference (ICI). The former is not a significant issue in today's cellular networks due to the use of orthogonal frequency-division multiple access (OFDMA) technology and base station controlled scheduling. On the other hand, ICI is a more relevant issue due to the emergence of small cells and the higher frequency reuse factor.

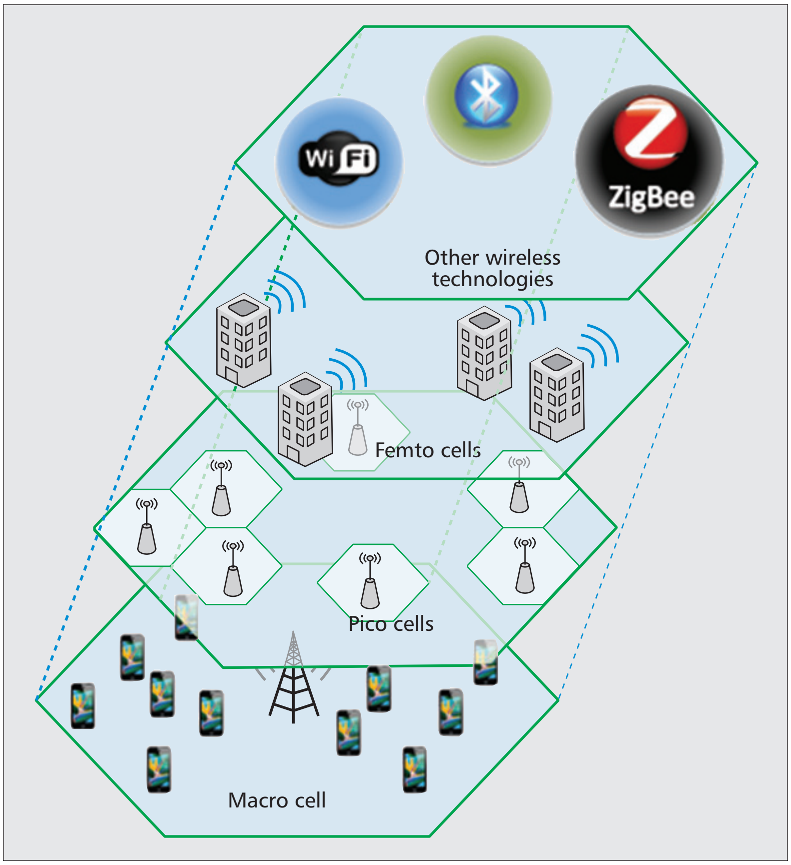


Figure 1. Today's wireless networks include multiple overlapping technologies.

Conventional cellular networks rely on the physical distance among cells and sectoring techniques to handle ICI. This approach cannot be used in dense deployments.

Spectral Efficiency — Due to high channel variation of wireless networks, the instantaneous channel quality of users varies significantly in a cell. Therefore, users experiencing low channel qualities can severely degrade spectrum efficiency. To counteract the impact of low channel quality users and to leverage the statistic fluctuations of channel qualities, opportunistic schedulers have been proposed and implemented, e.g., the Proportional Fair Scheduler or Max Rate [9]. A generic opportunistic scheduler always prioritizes the communication to the users with high channel quality and delays its communication to users with poor channel quality until their channel improves. Nevertheless, the scheduler waiting time should not result in transmission of expired data. Therefore, an opportunistic scheduler requires accurate information regarding the QoS constraints of the traffic and channel quality of the users. In a dense network, getting accurate feedbacks and processing them for a large number of users imposes high overhead in terms of data transmission and computational complexity. These problems are commonly tackled using selective feedbacks and machine learning techniques (to estimate the channel qualities). The latter reduces the transmission overhead but increases computational complexity.

[2] www.ict-crowd.eu

The studies show that D2D communication can potentially boost energy efficiency, throughput, delay, and fairness performances in cellular networks. To obtain even higher performance gain, some studies propose to integrate D2D communications and opportunistic scheduling in order to increase the spectral efficiency of the network.

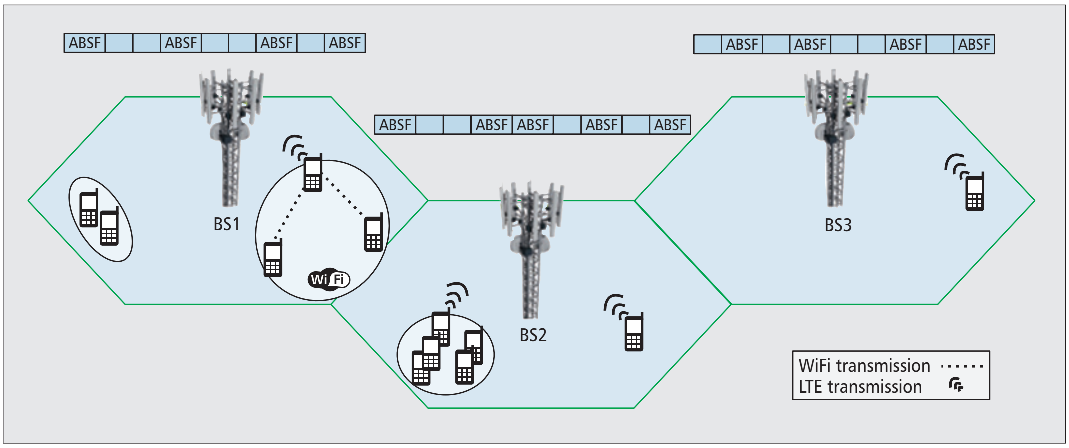


Figure 2. An example of a cellular network using D2D clustering and ABSF techniques.

Energy Efficiency — Cisco predicts that there will be more than 10 billion mobile devices by 2018, which makes the carbon footprint of wireless communication significant.[3] Notwithstanding the recent effort to improve the energy efficiency of wireless networks, the current infrastructure is not designed to be energy efficient. A popular method to reduce the power consumption of wireless networks is to put mobile devices and base stations to sleep whenever they are idle. Although the sleep functionality reduces the power consumption, it increases the delay in the network. This delay is proportional to the sleep period and the delay to switch from sleep to active mode.

Inflexible Infrastructure — The current cellular infrastructure is highly technology-dependent and inflexible to change. Any change in the standard operations of the network should go through the tedious standardization process in order to be finally implemented by the equipment manufacturers. As a result the solutions are implemented in the network with at least a few years delay. In fact, this is one of the reasons why the cellular technology could not catch up with the exponential network densification rate in wireless networks.

Feasible Solutions

Smart Interference Mitigation — As mentioned earlier, current ICI avoidance/control methods are no longer effective in today's dense wireless networks. It has been shown that although dense network deployments suffer more from ICI, the ICI power is not uniform over all radio frequencies. Therefore, ICI in dense networks is better managed if there exists a central controller with a bird's eye view of the occupied radio frequencies and ICI measurements. To this aim, researchers propose Intercell Interference Coordination (ICIC) techniques to take advantage of non-uniform ICI power distribution over the cell's radio spectrum. A very promising tool used to cope with the ICI problem is called ABSF. Specifically, ABSF allows the base stations to blank a set of subframes, which results in drastic ICI reduction. Note that the blank subframes can only be used for control signals, which is why those subframes are called *almost-blank*.

New Communication Paradigms — With the advent of cooperative communications, in particular D2D communications in cellular networks [10], researchers have started to probe the potential of this new paradigm [10]. In D2D communications, cellular users are allowed to communicate with each other without traversing the base station. The studies show that D2D communication can potentially boost energy efficiency, throughput, delay, and fairness performance in cellular networks. To obtain even higher performance gain, some studies propose to integrate D2D communications and opportunistic scheduling in order to increase the spectral efficiency of the network [7]. As mentioned before, opportunistic schedulers gain from the channel opportunistic scheduling of users with high channel quality and deferring the communication of those in low channel quality. It should be noted that this opportunistic gain is harnessed by the QoS requirements of the applications because the scheduler should schedule the user upon expiration of the QoS constraint even if the user has a poor channel quality. These D2D opportunistic schemes exploit the users with high channel quality to relay mobile traffic for those with lower channel quality. Therefore, the base station can transmit with higher Modulation Coding Scheme (MCS), which increases the spectral efficiency of the system.

Flexible Infrastructure — The above solutions and many other techniques for improving network performance in dense scenarios demand changes that are not foreseen by product manufacturers or by the standard. However, flexible architectures have been proposed, e.g. in international research programs like CROWD, and manufacturers are now endowing their devices with rich control interfaces. In this framework, SDN is an attractive paradigm that would allow network administrators to modify the behavior of the data plane by acting on the control plane. Although SDN was first proposed for wired networks [11], it is being considered as a viable solution to create a flexible wireless infrastructure.

[3] Cisco Visual Networking, "Global mobile data traffic forecast update, 2013-2018," White Paper, February, 2014.

Opportunistic Channel Utilization in Interference-Controlled Cells

We propose an SDN-controlled architecture for cellular networks with dense deployments of cells using a frequency reuse scheme. We assume mobile users have off-the-shelf dual radio devices (e.g. LTE and WiFi). To counteract the occurrence of interference and inefficient utilization of the licensed (and expensive) cellular spectrum, we propose to coordinate the activity of neighboring cells and to promote cooperation among users. The first component of our architecture is a mechanism that keeps inter-cell interference under control.We adopt ABSF for ICIC and use a smart and conservative approach to dynamically control and assign ABSF patterns to interfering base stations. The second component is a clustering technique that leverages D2D communications within a cell to enable fully opportunistic channel access without incurring fairness penalties.

ICI Mitigation Using ABSF

ABSF mitigates the inter-cell interference by assigning resources such that some base stations almost-blank their subframes, thus preventing their activity when the interference gets significant (see Fig. 2). Several solutions have already been proposed in the literature to leverage the ABSF mechanism. In our proposal, a central authority is in charge of acquiring the user channel conditions and then computing an optimal base station scheduling pattern, hereafter called the ABSF pattern, for the available subframes [12]. The algorithm exploits the ABSF technique to minimize the time required by the base stations to achieve high spectral efficiency when performing packet transmissions. The BSB algorithm provides a valid ABSF pattern by guaranteeing a minimum SINR for any user that might be scheduled in the system. Basically, the BSB algorithm tries to accommodate as many base stations as possible in the same scheduling interval, checking whether the minimum SINR constraint is fulfilled. In the case of a constraint violation, the algorithm sequentially removes the most interfering base stations, following the general guidelines provided for bin-packing heuristic methods. In addition, we slightly modified standard bin-packing procedures in order to accommodate base stations more than once within the ABSF pattern. However, the BSB algorithm is a conservative algorithm, which does not rely on the knowledge of base stations' scheduling decisions but just guarantees a decent user SINR, properly calculated a-priori.

D2D Clustering

D2D communication can occur over the cellular spectrum (i.e. inband) or the Industrial, Science and Medical (ISM) band (i.e. outband). Each approach has its pros and cons [10]. For instance, interference management is a major challenge in inband D2D communications because both cellular and D2D users share the same resources. While this type of interference is not an issue in outband D2D, the unregulated nature of the ISM band makes QoS guarantee a challenging task.

In CROWD, we have proposed a scheme, namely DRONEE (Dual-Radio Opportunistic Networking for Energy Efficiency), which leverages both D2D communications and opportunistic scheduling [7]. In DRONEE, neighboring mobile users can form a cluster using WiFi Direct (i.e. outband D2D) (see Fig. 2). In every cluster, only the cluster member with the highest channel quality (i.e. cluster head) will communicate with the base station (opportunistic scheduling). The cluster header is responsible for relaying the traffic of the other members to the base station. Using this scheme, the base station has a better chance of avoiding communicating with users with poor channel quality unless the whole cluster is suffering from deep fading and high interference. Moreover, the users with poor channel quality can enjoy higher transmission rates by relaying through the cluster head. Simulation results show that the D2D clustering scheme can achieve significant throughput, energy efficiency, and fairness gain in comparison to conventional cellular networks. For more details on D2D-enabled opportunistic clustering schemes, please refer to [7].

Interoperation of ABSF Control and D2D-Enabled Opportunistic Clustering Techniques

The control of ABSF patterns is key to orchestrating the activity of multiple cells, while the D2D-based opportunistic channel access is key to using the radio spectrum efficiently. The combination of the two mechanisms is beneficial for the performance of the networks, as we will show in Section IV. Moreover, it is important to note that the two mechanisms help each other to achieve their goals, therefore creating a positive feedback effect. On one hand, the reason why D2D clusters with opportunistic scheduling benefits from the presence of inter-cell interference relies on the reduction of uncontrolled interference entering a cell from the outside. Therefore, intra-cell channels are affected by less unpredictable interference and are more stable, which means that opportunistic changes of cluster heads will occur less often. On the other hand, the computation of ABSF patterns via the BSB algorithm is simplified in the presence of clusters in the cells, since the algorithm will only need to consider clusters instead of users. As a consequence, the complexity of the BSB reduces with a cubic function of the average cluster size. Moreover, the fact that few clusters (which with high probability have a cluster head experiencing a good channel) are scheduled instead of many regular users (many of which with poor channels) reduces the number of cases in which BSB assigns very conservative ABSF patterns. Those conservative patterns are typically due to the presence of poor users, although the probability that the base station will schedule those users is not known to the controller.

SDN

ABSF and D2D clustering schemes both require a wireless infrastructure with high interoperability among heterogeneous networks. This can be easily achieved using the SDN architecture pro-

The control of ABSF patterns is key to orchestrate the activity of multiple cells, while the D2D-based opportunistic channel access is key to use the radio spectrum efficiently. The combination of the two mechanisms is beneficial for the performance of the networks.

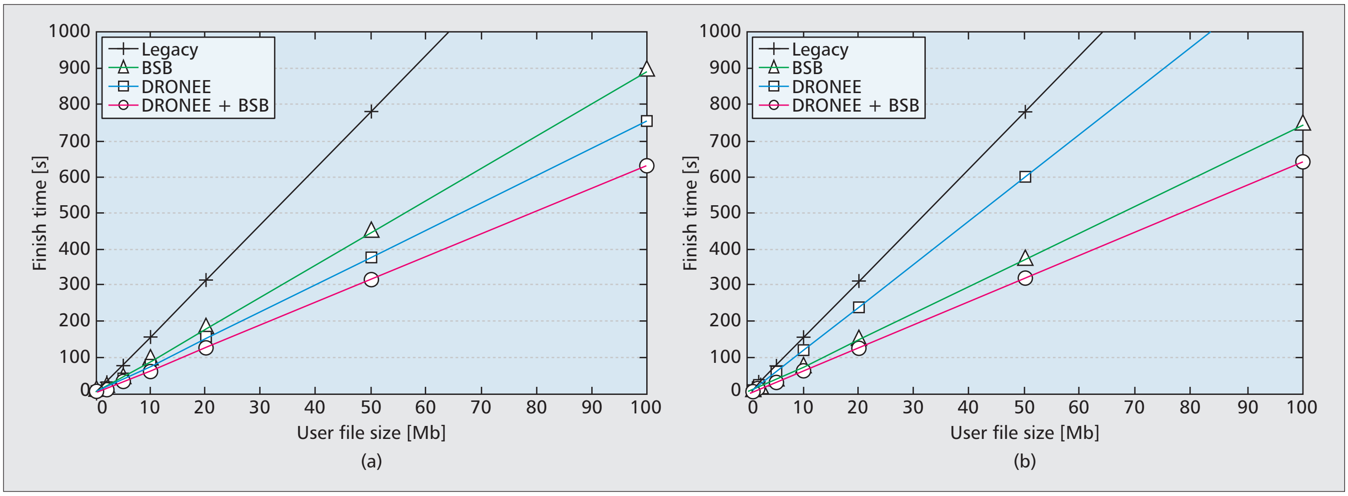


Figure 3. Finish time in a network with 10 base stations, 500 users, and maximum cluster size for DRONEE equal to 10 users: a) Inter-site distance equal to 300 m; and b) inter-site distance equal to 30 m.

posed in the CROWD project. In fact, the proposed SDN-based architecture offers a flexible network with the capability to accommodate other wireless technologies. Therefore, the CROWD SDN solution allows the integration of ABSF and D2D clustering into existing cellular architectures, and it paves the way for future enhancing proposals that may require unforeseen infrastructural modification.

In the next section our solutions are validated through simulation and benchmarked against legacy network operation schemes. We assume that the practical modifications required for the implementation of our solutions are supported by the CROWD SDN architecture. Note that SDN is not a necessity but a suitable tool that offers flexible and upgradable solutions. In addition, SDN enables the coordination of intra-cell and inter-cell resource optimization mechanisms, e.g. by creating per-cluster statistics (based on user statistics) to be passed to the BSB algorithm.

Evaluation

Let us compare the performance of BSB and DRONEE, and their compound impact on the performance of dense cellular networks. We compare the performance achieved with our proposed schemes to the ones achieved in conventional cellular networks without ICIC. We evaluate our solution for an LTE network with 20 MHz bandwidth. Users in this network are randomly distributed over the coverage area of the cells according to a uniform distribution. Each user is trying to download a 1 Mb file and each simulation terminates when all users have completed their download (we call this the *finish time*). Users are allowed to form clusters and use D2D within 50 m radius. The cluster formation is done using the *merge and split* algorithm [13] and each simulation is repeated 20 times.

Figures 3a and 3b show the finish time achieved under the adoption of different schemes in a network with 500 users and 10 base stations regularly distributed over the simulated area with an inter-site distance of 300 m. For D2D-based clustering, base stations adopt a weighted round robin policy and assign resources to cluster heads with weights proportional to the cluster sizes. In both figures we can see that BSB and DRONEE (with clusters of at most 10 users) significantly reduce the finish time in comparison to conventional cellular networks. This improvement is due to efficient ICI management of BSB and high spectral efficiency of DRONEE. In Fig. 3b we reduced the intersite distance by a factor of 10, which results in increased ICI. As expected, here BSB shows better performance because it is designed to deal with high ICI. Interestingly, the combination of BSB and DRONEE demonstrates even higher gain. This happens because these two schemes aim to solve two different issues in dense network, so they are complementary. However, the gain stemming from the coupled control of BSB and DRONEE is not equivalent to the sum of the gain from BSB and DRONEE clustering because both techniques rely on wireless channel diversity to improve the performance. Therefore, the diversity gain of clustering after applying ICIC via ABSF patterns with BSB is lower because channel quality of users increases due to lower interference (i.e. the opportunistic scheduling gain is lower).

Figure 4 illustrates the impact of user density on the finish time where the maximum cluster size is 5. The finish time of all schemes follows an increasing trend with an increasing number of users, due to increased traffic load and interference. However, the network densification is better handled by jointly coordinating BSB and DRONEE because they are designed to reduce inter-cell and intra-cell inefficiencies. Moreover, DRONEE facilitates the interference control operated by BSB and BSB facilitates the management of clusters in DRONEE, although this aspect cannot be shown in the figure. Clustering schemes such as DRONEE take advantage of the user density to form more clusters, which results in better opportunistic gain and lower interference. The BSB algorithm orchestrates interference and allows base stations to allocate more traffic in less subframes. The results shown in the figure also confirm that the compound impact of inter-cell and intra-cell resource allocations through BSB and DRONEE reduces the finish time drastically (60% less, w.r.t. legacy network operation).

The impact of cluster size on the finish time is shown in Fig. 5. Performance improves as the maximum allowable cluster size increases because the opportunistic clustering gain increases with the cluster size [7]. In the figure, we show the average finish time of all the users (solid lines) and the finish time of the users in clusters with exact size of n users, where n is the value reported in the horizontal axis of the figure (dotted lines). The latter shows a stable decreasing trend while the former has higher variation because smaller clusters and single unclustered users can take much longer to finish their download. The figure also shows that the clustering gain saturates for clusters bigger than 15. This is due to the fact that the room for improving the aggregated channel quality of the cluster becomes marginal in big clusters.

Discussion

The evaluation results confirm the great performance gain of ABSF and D2D clustering schemes, in particular for BSB and DRONEE and their combination. The gain is evident with respect to legacy-operated networks, and exemplifies the capability of the proposed schemes to leverage network densification as a resource. The CROWD approach is advantageous because it proposes SDN to combine DRONEE and BSB, and easily embed them into today's cellular network architecture. Moreover, it has been shown that D2D clustering can be readily integrated into the LTE-A infrastructure with minimal modifications [14]. Therefore, with the current capabilities of WiFi Direct and LTE-A, D2D clustering is no longer a far-fetched concept. Moreover, ABSF is already available in LTE-A, and our proposed algorithm can be readily implemented in the current system.

The merits of our CROWD solution are not limited to throughput increment. For instance, D2D clustering enhances energy efficiency by allowing mobiles to switch to a low power consumption technology (i.e. WiFi) and to reduce the overall transmission time because only the users with the highest channel quality communicate with the base station. Cluster formation also paves the way toward improving user fairness in the system. Once a cluster is formed, a virtual pool of cellular resources can be created that is equivalent to the aggregate of the individual resources of each cluster member. Base stations can exploit these virtual pools to use cluster heads to provide more data to the users with lower channel quality and avoid the starvation issue well known for opportunistic schedulers. Our evaluation results indicate that bigger cluster sizes result in higher gain. However, the relation between cluster size and gain is not linear and most of the gain is achieved when the cluster size is between 5 to 10, which happens to be small enough to avoid overwhelming signaling and contention overhead in WiFi D2D operation.

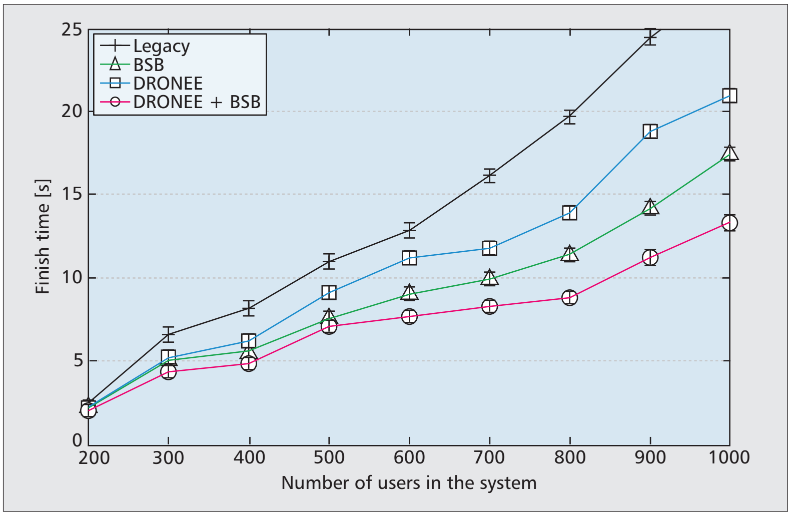


Figure 4. The impact of user density on finish time with 10 base stations and inter-site distance equal to 300 m (results for DRONEE and DRONEE+ BSB are achieved with clusters of max five users).

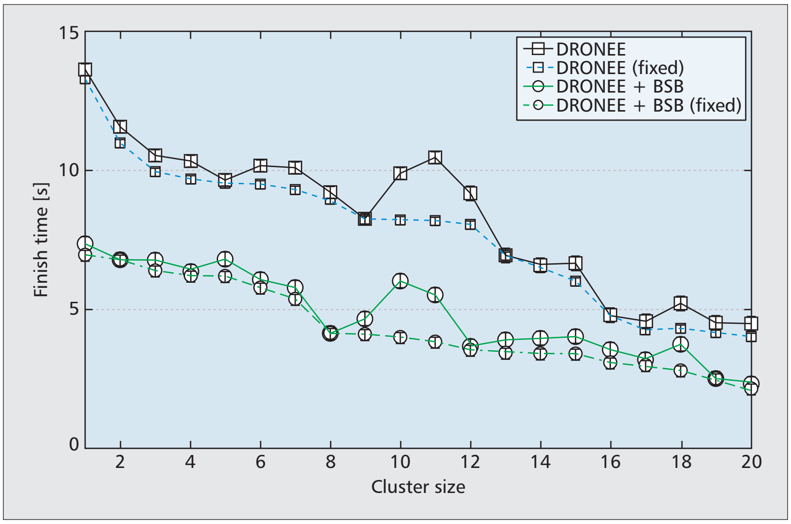


Figure 5. The impact of cluster size on finish time with 10 base stations and inter-site distance equal to 300 m (curves labeled with (Fixed) correspond to the statistics obtained with clusters of exactly n users, where n is the value reported in the horizontal axis).

Conclusions

In this article we have shown that the interference arising in dense wireless networks can be counteracted by controlling inter-cell interference while using intra-cell resources opportunistically. Specifically, we have shown the compound beneficial impact of BSB (a mechanism proposed for inter-cell resource allocation) and DRONEE (a mechanism proposed for channel opportunistic use of cellular resources). Our results showed that ICIC can be implemented via smart allocation of ABSF patterns for interfering base stations, and most importantly the impact of ICIC can be magnified by adopting channel opportunistic scheduling within the cells. Indeed, D2D communications and clustering techniques not only improve the spectral efficiency within the cell, but also reduce the complexity of ICIC algorithms such as BSB. The proposed joint orchestration of BSB and DRONEE represents a powerful and feasible solution for extremely dense wireless networks, and can be suitably implemented by means of SDN controllers.

The proposed joint orchestration of BSB and DRONEE represents a powerful and feasible solution for extremely dense wireless networks, and can be suitably implemented by means of SDN controllers.

Acknowledgements

The research leading to these results was supported by the CROWD project, under the European Union's Seventh Framework Programme (grant agreement no. 318115).

References

[1] N. Bhushan *et al.*, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 82–89.

[2] R. Bendlin *et al.*, "Fast Distributed Multi-Cell Scheduling with Delayed Limited-Capacity Backhaul Links," *IEEE ICC*, June 2009.

[3] M. Rahman and H. Yanikomeroglu, "Multicell Downlink OFDM Subchannel Allocations Using Dynamic Inter-Cell Coordination," *IEEE GLOBECOM*, 2007, pp. 5220–25.

[4] —, "Enhancing Cell-Edge Performance: A Downlink Dynamic Interference Avoidance Scheme with Inter-Cell Coordination," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, 2010, pp. 1414– 25.

[5] M. C. Necker, "Interference Coordination in Cellular OFDMA Networks," *IEEE Network*, vol. 22, no. 6, Dec. 2008, p. 12.

[6] V. Sciancalepore, V. Mancuso, and A. Banchs, "Basics: Scheduling Base Stations to Mitigate Interferences in Cellular Networks," *IEEE WoWMoM*, June 2013, pp. 1–9.

[7] A. Asadi and V. Mancuso, "Dronee: Dual-Radio Opportunistic Networking for Energy Efficiency," *Computer Commun.*, 2014.

[8] Q. Wang and B. Rengarajan, "Recouping Opportunistic Gain in Dense Base Station Layouts through Energy-Aware User Cooperation," *IEEE WoWMoM*, 2013, pp. 1–9.

[9] A. Asadi and V. Mancuso, "A Survey on Opportunistic Scheduling in Wireless Communications," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 4, 4th quarter 2013, pp. 1671–88.

[10] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Commun. Surveys & Tutorials*, 2014.

[11] M. Yang *et al.*, "Openran: A Software-Defined Ran Architecture Via Virtualization," *ACM SIGCOMM*, 2013, pp. 549–50.

[12] V. Sciancalepore *et al.*, "Interference Coordination Strategies for Content Update Dissemination in LTE-A," *IEEE INFOCOM*, 2014.

[13] W. Saad *et al.*, "Coalitional Game Theory for Communication Networks," *IEEE Signal Proc. Mag.*, 2009.

[14] A. Asadi and V. Mancuso, "WiFi Direct and LTE D2D in action," *IFIP Wireless Days*, Nov 2013, pp. 1–8.

Biographies

ARASH ASADI (arash.asadi83@gmail.com) received a master's degree with the highest distinction in telematics engineering from the University Carlos III of Madrid, and a master's degree in telecommunication engineering from Multimedia University. He is currently with IMDEA Networks Institute pursuing his Ph.D. studies. His primary research interests are resource allocation in wireless networks, cooperative networks, and D2D communications.

VINCENZO SCIANCALEPORE (vincenzo.sciancalepore@imdea.org) received a B.Sc. in computer engineering from the University Politecnico di Bari in Italy in 2008. In 2011 he received his M.Sc. degree in telecommunications engineering from the University Politecnico di Milano. In 2012 he received a M.Sc. degree in telematics engineering from the University Carlos III of Madrid. Currently he is working as a Ph.D. student in a joint Ph.D. program between the Institute IMDEA Networks in Madrid and Politecnico di Milano, focusing his activity on inter-cell coordinated scheduling for LTE-Advanced networks and device-to-device communication.

VINCENZO MANCUSO (vincenzo.mancuso@imdea.org) has been a research assistant professor at IMDEA Networks Institute (Madrid, Spain) since September 2010. He has build his research experience by working with the University of Palermo (Italy), from which he received a Ph.D. in electronic, computer science, and telecommunications in 2005, Rice University (Houston, TX, USA), and INRIA Sophia Antipolis (France). His research activities focus on analysis, design, optimization and experimental evaluation of protocols and architectures for efficient wireless networks.

Toward 5G DenseNets: Architectural Advances for Effective Machine-Type Communications over Femtocells

Massimo Condoluci, Mischa Dohler, Giuseppe Araniti, Antonella Molinaro, and Kan Zheng

Massimo Condoluci, Giuseppe Araniti, and Antonella Molinaro are with University Mediterranea of Reggio Calabria.

Mischa Dohler is with King's College London and Worldsensing.

Kan Zheng is with Beijing University of Posts and Telecommunications.

ABSTRACT

Ubiquitous, reliable and low-latency machine-type communication, MTC, systems are considered to be value-adds of emerging 5G cellular networks. To meet the technical and economical requirements for exponentially growing MTC traffic, we advocate the use of small cells to handle the massive and dense MTC rollout. We introduce a novel 3GPP-compliant architecture that absorbs the MTC traffic via home evolved NodeBs, allowing us to significantly reduce congestion and overloading of radio access and core networks. A major design challenge has been to deal with the interference to human-type traffic and the large degree of freedom of the system, due to the unplanned deployments of small cells and the enormous amount of MTC devices. Simulation results in terms of MTC access delay, energy consumption, and delivery rate corroborate the superiority of the proposed working architecture.

INTRODUCTION

The support of machine-type communication (MTC) via evolved Long Term Evolution-Advanced (LTE-A) [1] represents one of the main growing challenges for cellular network providers in order to fulfill the requirements of future fifth generation (5G) wireless networks [2]. MTC has promising economic and strategic value in the mobile market scenario since an exponential growth in the data traffic generated by heterogeneous devices (e.g., smart meters, signboards, cameras, remote sensors) is expected with consequent unprecedented opportunities and business models for telco operators in different fields (e.g., transport and logistics, smart power grids, e-health, home and/or remote surveillance) [3].

MTC represents a novel transmission paradigm where devices send their data to remote servers or to other machines (e.g., actuators) without human intervention. MTC has a set of unique and challenging characteristics [4] (e.g., group-based communications, low or no mobility, time-controlled, time-tolerant, secure connection) which require technically advanced solutions currently under investigation by academia, industry, and standards bodies, such as the Third Generation Partnership Project (3GPP) [5, 6].

Due to the high (and unpredictable) number of MTC devices expected to simultaneously access the cellular network [7], **congestion and overloading of radio access and core networks**[1] are the prime issues to be solved in order to guarantee low-latency and low-energy MTC, and to minimize the impact of MTC on these network segments. Another important issue is related to the observation that many machines are geographically **located in a very confined and coverage-limited area** (e.g., sensors/actuators in a hospital or a refinery). As a consequence, the radio access network beyond LTE-A should be able to efficiently manage several hotspots, many of which might be located in challenging positions (e.g., indoors or at the cell edge). AThe above mentioned issues are exacerbated by the strict requirements on the design of MTC terminals, which should be low cost (i.e., low complexity and computational capabilities) and have **low energy consumption** in order to guarantee extended lifetimes. Furthermore, the M2M solution provider may not want to be **dependent on the coverage and rollout strategy of a given operator** and may thus want to provide coverage on its own. Finally, another design driver is that **MTC should not affect the performance of traditional human-type communication (HTC)** in the form of, say, voice/data calls [4].

To the best of our knowledge, the architecture outlined in this work is *the first of its kind that meets the above design criteria by explicitly uniting MTC and femtocells*. We show the architectural enhancements able to adequately support the extremely dense MTC deployment scenarios through the exploitation of femtocells. The proposed system architecture guarantees low-latency MTC without meaningful additional costs compared to non-3GPP wireless networks and without affecting the performance of HTC traffic, as shown later. The exceptional performance under ultra-dense MTC rollouts is demonstrated. Finally, we summarize our findings and outline some future work.

[1] Congestion arises when a base station does not have enough resources to support the traffic load generated by the enormous number of MTC devices within the cell, while overloading is due to very large loads at the network entities that have to manage data/control traffic from/to MTC devices. Such issues increase the time needed by MTC devices to access the network and transmit data.

0163-6804/15/$25.00 © 2015 IEEE

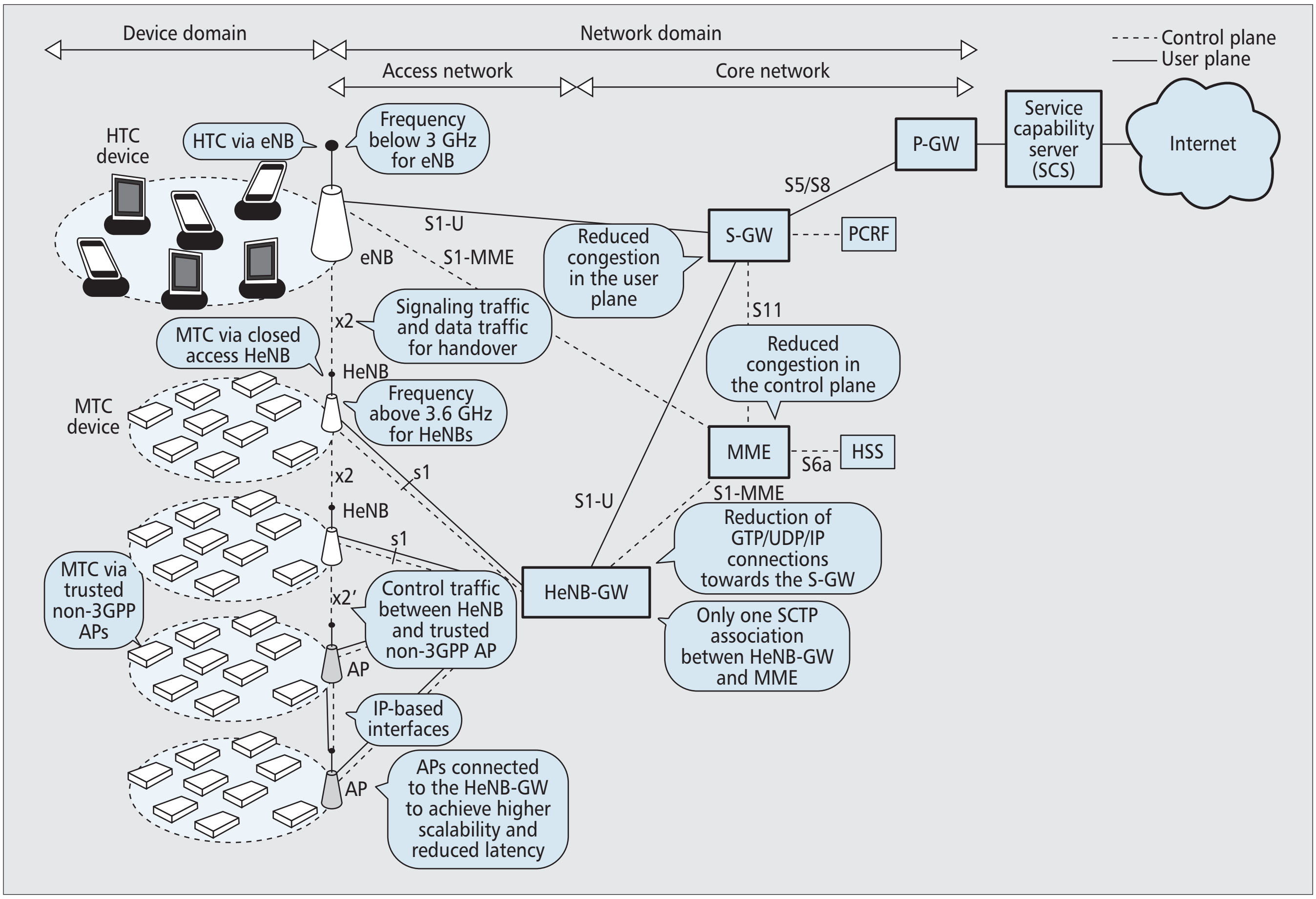

Device domain
Network domain
Access network
Core network
Control plane
User plane
HTC device
HTC via eNB
Frequency below 3 GHz for eNB
eNB
S1-U
S1-MME
Reduced congestion in the user plane
S-GW
PCRF
P-GW
S5/S8
Service capability server (SCS)
Internet
x2
Signaling traffic and data traffic for handover
S11
Reduced congestion in the control plane
MTC via closed access HeNB
HeNB
MTC device
Frequency above 3.6 GHz for HeNBs
MME
HSS
S6a
x2
s1
HeNB
S1-U
S1-MME
s1
Reduction of GTP/UDP/IP connections towards the S-GW
MTC via trusted non-3GPP APs
x2'
Control traffic between HeNB and trusted non-3GPP AP
HeNB-GW
Only one SCTP association betwen HeNB-GW and MME
AP
IP-based interfaces
AP
APs connected to the HeNB-GW to achieve higher scalability and reduced latency

Figure 1. Enhanced network architecture for ultra-dense MTC access to the 3GPP LTE-A core via HeNBs/HeNB-GWs and trusted non-3GPP APs.

Scalable Network Architecture for Ultra-Dense MTC

We consider a scenario where MTC is handled by a network operator that provides control (e.g., subscriber management) and data functionalities (e.g., traffic transport) via a 3GPP LTE-A system. According to the European Telecommunications Standards Institute (ETSI) [8], the MTC application is typically hosted by an application server (AS), which may be directly connected to the operator network or make use of a services capability server (SCS) that offers additional control (e.g., device triggering) and data services for MTC. The SCS may be controlled by the MTC service provider or the 3GPP network operator. Also, a hybrid solution is permitted, where the user plane communications with the MTC device is directly managed by the AS, and the control plane connections are handled by the SCS.

The high number of machines (mainly located in challenging positions within the cell) that simultaneously access the network poses several issues for traditional macrocellular systems, which are not able to guarantee the requirements of ubiquitous connectivity and high capacity of MTC. An effective transport network able to support MTC has to efficiently manage the expected extremely dense scenarios, and improve both radio access and core networks by reducing traffic overload and interference with the HTC traffic, while guaranteeing low-latency interconnection with non-3GPP networks.

The proposed evolved 3GPP architecture tailored to efficiently support MTC is illustrated in Figure 1. In our architecture, LTE-capable MTC devices communicate directly with home-evolved NodeBs (HeNBs), which are low-cost femtocells (with economic efforts close to those of non-3GPP access points in terms of device cost, installation, and manteinance) for local-area access with low-power transmission (less than 100 mW) capability [9]. Non-3GPP devices access through trusted non-3GPP access points (APs) that are operator controlled and managed by the HeNB gateway (HeNB-GW) like legacy HeNBs. The HeNB-GW aggregates traffic from a large number of HeNBs and trusted non-3GPP APs into the existing core network. Motivations and advantages of the illustrated architectural choices are individually analyzed in the following.

MTC and HeNB Network Architecture

The exploitation of HeNBs for MTC, instead of macro-eNodeBs (MeNBs), in the radio access network achieves the following goals.

Stronger separation of MTC and HTC traffic: HeNBs may be installed by cellular customers in the form of individuals and companies/industries aiming to (mainly but not exclusively

With the aim to corroborate the expected MTC enhancements of the proposed architecture, a simulation campaign has been carried out through a 3GPP-calibrated system-level simulator. The analysis focuses on the radio access network, which is the most solicited segment by the multitude of MTC devices.

— see later section) interconnecting their own MTC devices in specific areas, such as in hospitals, offices, laboratories, and plants. Once plugged in, HeNBs connect to the 3GPP core network via the digital subscriber line/fiber and represent the entry points for LTE-A MTC devices. MeNBs are instead exploited by the operator for HTC services without being affected by the MTC traffic load.

Closed access: MTC could be handled advantageously via closed access HeNBs. According to this solution, access to HeNB(s) is allowed only to those machines that belong to a given closed subscriber group (CSG) [9]; for instance, a customer can admit only its own devices through its own HeNB. Closed access HeNBs offer the customer/operator *secure* access and the possibility to perform *load balancing* through CSGs creation and management.

Coverage extension: HeNBs are useful to provide local connectivity for MTC devices located in remote or challenging locations (e.g., rural areas, indoor deployments, smart meters in the basements of the buildings) *without* requiring network replanning.

The simultaneous exploitation of MeNBs and HeNBs requires effective solutions to reduce or avoid intercell interference. While co-channel deployments have been advocated (and well researched) in the past, an interesting study [2] recently stated that *frequency-separated* deployments are *the most promising solution* addressed by telco operators. Licensed spectrum below 3 GHz has already been widely exploited for cellular services via MeNBs; however, additional higher frequency bands (e.g., 3.5 GHz and above) have recently become available. Such frequencies are challenging for macrocell deployments due to propagation characteristics, while they are suitable for HeNBs communicating over relatively short range. This proposed frequency-separated deployment avoids intercell interference and relaxes the radio frequency (RF) requirements of MTC devices with a consequent reduction of equipment costs [2].

In the proposed network architecture, the *X2* interface handles the exchange of control traffic between MeNBs and HeNBs for system parameters configuration/reconfiguration (e.g., frequency band selection). The X2 interface can be further exploited as a low-latency interface to exchange data traffic for time-critical events such as handovers [10].

Reducing the Core Network Load

Femtocells are attached via digital subscriber lines/fiber to the 3GPP LTE-A core network (a.k.a. system architecture evolution, SAE), which is composed of two entities [11]: the mobility management entity (MME) and the serving gateway (S-GW). The MME is a control plane unit that manages security functions (e.g., authentication and authorization) and handovers and handles idle mode state and CSG subscription. The S-GW works in the user plane by routing and forwarding data packets to and from the MeNBs/HeNBs. In addition, it offers connectivity to external networks through the packet data network gateway (P-GW).

The 3GPP has proposed solutions for both direct and non-direct connection of HeNBs to core network entities. Although the former solution is attractive to reduce the number of network units needed for HeNB deployment, it suffers from several inefficiencies in terms of MME/S-GW overload when the number of HeNBs as well as the number of connections per HeNB are large. In the proposed architecture, the HeNB-GW — part of the backhaul and securely connected to the core — has the key role of being a concentrator of several HeNBs for both control and user planes. This has obvious advantages in terms of scalability and load reduction in the core network. The benefits offered by the mandatory use of an HeNB-GW can be summarized as follows.

One Stream Control Transmission Protocol (SCTP) association between the HeNB-GW and MME: In LTE-A, an SCTP association is created between two entities exchanging control plane traffic. Through the use of the HeNB-GW, the presence of a large number of HeNBs does not imply congestion at the MME. Indeed, only the HeNB-GW transmits SCTP heartbeat messages to the MME instead of each single HeNB. Furthermore, the number of SCTP association establishments and releases due to HeNBs switching on/off is minimized.

S-GW scalability: The number of GPRS Tunneling Protocol (GTP), UDP, and IP connections between the HeNB-GW and S-GW are drastically reduced compared to a direct HeNB-to-HeNB connection. In this way, the number of HeNBs may increase without an increase in the number of UDP/IP paths and GTP Echo messages managed by the S-GW.

Paging optimization: Optimized paging mechanisms for downlink data transmission to the managed HeNBs can be implemented within the HeNB-GW to reduce latency.

According to the architectural solution proposed in Fig. 1, two challenges need to be evaluated. The first one is due to the fact that the HeNB-GW has to switch from the HeNB-GW–S-GW tunnel to the HeNB-GW–HeNB tunnel (and vice versa). Consequently, the higher the traffic from the machines, the higher the load at the HeNB-GW in the user plane. Nevertheless, it is worth noting that the HeNB-GW is the only entity overloaded when the traffic from the machines increases, and this influences only the performance of devices connected to the femtocells managed by the HeNB-GW. On the contrary, by using alternative solutions where the femtocells are directly connected to the S-GW, an increase in the MTC traffic involves overloading at the S-GW with influence on the traffic from/to all the nodes (i.e., a large set of macrocells and femtocells) managed by the S-GW. The second issue is that the HeNB connects to a single HeNB-GW at one time, and this reduces the redundancy and load sharing possibilities in comparison with other architectural variants. It is worth noting that we foresee the use of the X2 interface among different femtocells. Such an interface can be further exploited as a low-latency interface to increase the redundancy as well as sharing possibilities. As a consequence, it is clear that our proposal is

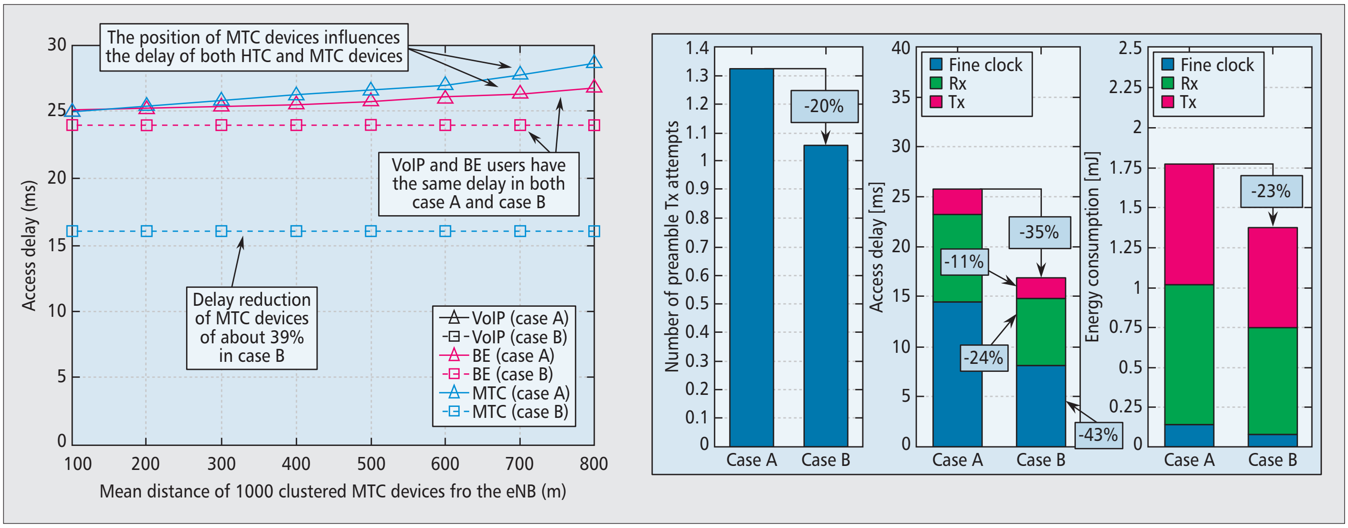


Figure 2. Access delay for HTC and MTC terminals vs. distance of 1000 MTC devices from the MeNB (left) and comparison between RA via MeNB and HeNB when devices are 100 m from the MeNB (right).

more effective to reduce the overall system overload than other solutions.

Interconnection with Non-3GPP APs

An important issue to consider is the interconnection of machines belonging to non-3GPP networks, such as IEEE 802.11ah (low-power Wi-Fi), which is gaining in popularity. In the proposed network architecture, we foresee *trusted* non-3GPP APs [12], which are typically operator-managed Wi-Fi APs that provide access functionalities, over-the-air encryption, and secure authentication and billing functionalities.

The communication between Wi-Fi-enabled devices and LTE-A MTC terminals can be achieved, for instance, through traditional Internet communication. In this case, the data sent by the Wi-Fi machine is conveyed to the LTE-A terminal via the P-GW and S-GW. This solution may involve intolerable data transmission delay, especially if two MTC devices are geographically close (e.g., in the same room).

To cope with this issue, we propose the introduction of a short data path (while other functionalities, e.g., access, authentication and billing, are the same as in [12]) by connecting the trusted non-3GPP APs to the HeNB-GW. The APs exchange data through the non-3GPP interface with the served MTC devices, while they appear like HeNBs to the HeNB-GW, with the following advantages:

- **Reduced latency**, since data communication between non-3GPP and LTE-A devices can be handled through the HeNB-GW
- **Additional scalability**, since the increase in the number of non-3GPP APs does not overload the S-GW

In addition, the latency of some control procedures can be cut by reducing the amount of operations in the core network entities. Indeed, since APs and HeNBs share the same protocol stack, a novel logical IP-based control interface, X2′ in Figure 1, ought to be standardized to support system parameters configuration and for the handover of those devices equipped with both non-3GPP and LTE-A interfaces .

Analysis of MTC Radio Access and Data Transmission

With the aim of corroborating the expected MTC enhancements of the proposed architecture, a simulation campaign has been carried out through a 3GPP-calibrated system-level simulator. The analysis focuses on the radio access network, which is the most solicited segment by the multitude of MTC devices.

The cell layout, radio channel model, and power transmission levels are set according to the 3GPP macrocell case #3 [13]. HTC and MTC devices are in the coverage area of one macrocell site (hexagonal grid, 3 sectors per site, 2 GHz carrier frequency, 5 MHz channel bandwidth), while 18 neighboring cells are considered as interfering cells (the inter-site distance is set to 1732 m). We considered 30 VoIP and 20 active best effort (BE) users which are uniformly distributed in the macrocell, while 1000 MTC devices are clustered in a restricted area (with a radius equal to 50 m) of the cell.

In order to effectively evaluate the benefits introduced by the use of HeNBs in the radio access network, we compare two scenarios. The first one (case A) is a macro area environment, where both HTC and MTC devices are attached to the MeNB. In the second scenario (case B), HTC users are attached to the MeNB while all MTC is handled via the HeNB. To further consider the impact of the position of MTC devices on HTC communications, we varied the mean distance of MTC terminals from the MeNB.

Random Access Procedure

The random access (RA) procedure [14] is performed by a device in several cases; for example, upon initial network access, in idle mode, during handover, and for connection re-establishment after a radio link failure. Contention-based access is managed via a four-message handshake between the device and the base station. A well designed network should be able to guarantee low-latency access to MTC devices. This aspect is especially crucial, considering that commonly MTC devices

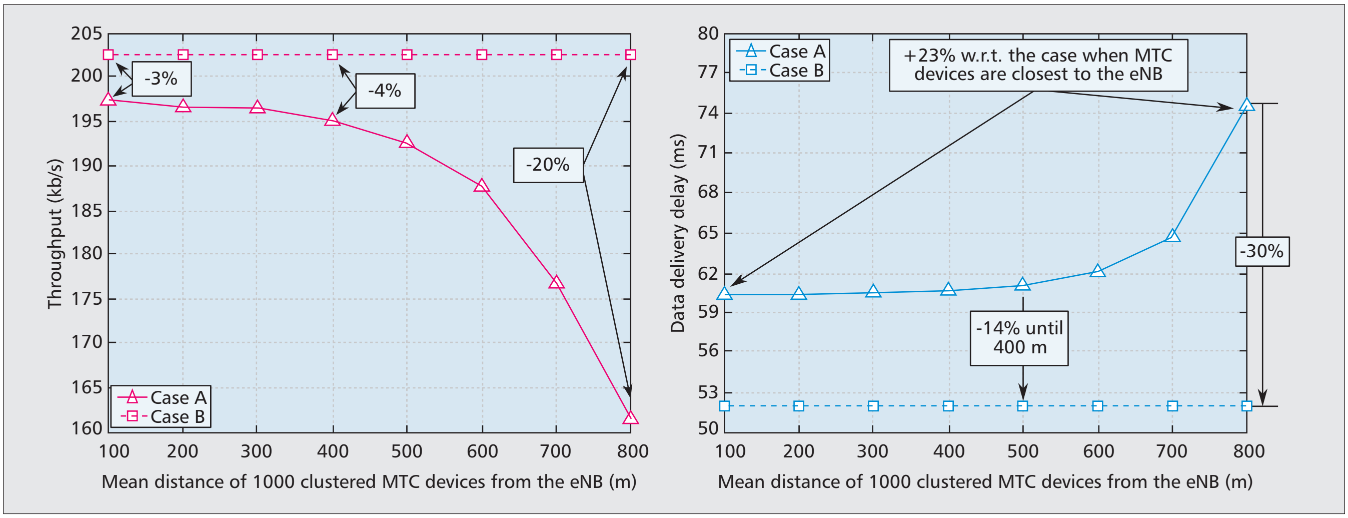


Figure 3. The impact of MTC on throughput of BE users (left) and latency data transmission of MTC devices (right).

transmit only one data packet after succeeding with the RA procedure. As a consequence, reducing RA latency allows the control overhead necessary for data transmission to be decreased.

To analyze this aspect, we considered a period of 60 s during which MTC devices and HTC users perform one RA procedure, and we measured the delay spent for such a procedure (the *access delay*). The arrival rates of HTC and MTC terminals are uniformly distributed within the considered period of 60 s. The RA parameters are set according to [15].

The achieved results are shown in Fig. 2. Considering case A, when MTC devices are in challenging positions within the cell (i.e., larger distances from the MeNB), both MTC and HTC users experience higher RA delays. It emerges, instead, that the use of HeNB (i.e., case B) for MTC traffic has the positive effect of reducing the access delay of both HTC and MTC terminals; the main benefit is for MTC devices, which achieve an access delay reduction of about 39 percent in the considered scenario, while HTC users gain less than 3 ms. It is worth noticing that the MTC access delay is insensitive to the position of MTC devices in the cell with respect to the MeNB, since the HeNB is always located close to MTC devices.

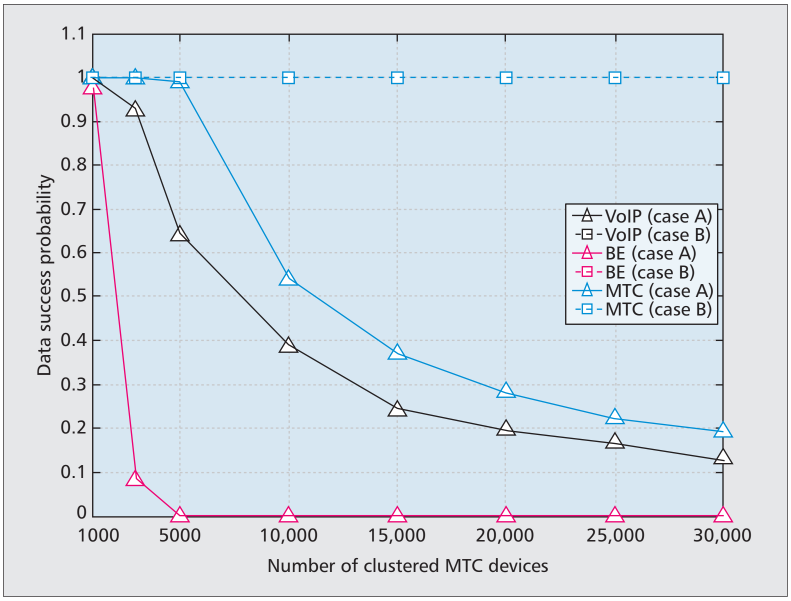
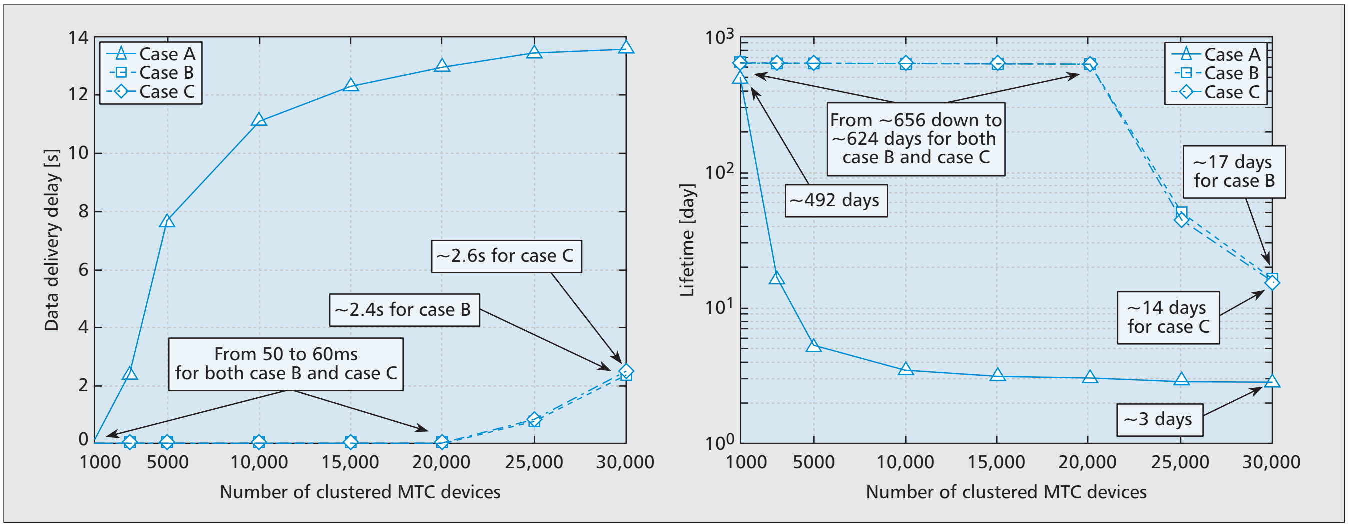


Figure 4. Data success probability for HTC and MTC devices by varying the number of clustered MTC devices located at the cell edge.

The MTC behavior is further analyzed on the right of Fig. 2 and confirms the enhancement introduced in the RA procedure by use of the HeNB. It is worth noticing that in case B, the number of preamble transmissions (i.e., the first message of an RA procedure) is reduced by a factor equal to 20 percent compared to the case when MTC terminals are attached to the MeNB (i.e., case A). This is not due to a lower preamble collision rate (which is the same for both cases and depends on the number of machines that access in the same RA slot), but due to the better channel conditions experienced by MTC devices due to the shorter device/base station distance, which increases the probability of successful preamble transmission. By reducing the number of RA procedures necessary to accomplish radio access, the MTC devices spend less time in *fine clock* (i.e., the waiting time for uplink transmission or downlink reception), *receiver*, and *transmission* states. A further positive effect is the decrease of energy consumption in the MTC devices by 23 percent.

Data Transmission

Data transmission from HTC and MTC devices poses several issues since different requirements for HTC and MTC services need to be considered. VoIP communication can be considered as a *periodic* transmission (20 bytes every 20 ms during on periods) and is handled by the MeNB through semi-persistent scheduling. The BE traffic is instead scheduled according to the resources still available after the transmission of flows with higher priority. Finally, MTC devices usually transmit only one small message after the radio access procedure, thus asking for few system resources. The main concern related to MTC for the system is instead due to the massive number of instantaneous transmitting MTC



~2.6s for case C

~2.4s for case B

From 50 to 60ms for both case B and case C

From ~656 down to ~624 days for both case B and case C

~492 days

~17 days for case B

~14 days for case C

~3 days

Figure 5. Delivery time (left) and lifetime (right) of MTC devices.

devices. We considered a period of 60 s during which MTC and HTC terminals perform data transmission. MTC devices are assumed to transmit one 200-byte message, while BE users are served by the MeNB through a maximum throughput scheduler. Finally, VoIP users are considered as "background" services since they are semi-persistently scheduled by the MeNB.

The transmission performance is shown in Fig. 3. The mean throughput of BE users is shown on the left. When MTC devices are served via the MeNB (i.e., case A), the throughput of BE users decreases, more so at greater distances. This is due to the fact that the greater the distance from the MeNB, the poorer the channel quality of MTC devices. As a consequence, they require a higher portion of system resources for data transmission, and this causes a throughput reduction for BE users. When MTC devices are handled by the HeNB (i.e., case B), the throughput of BE flows increases by about 4 percent until the case when MTC devices are 400 m from the MeNB; then it increases by a factor equal to 20 percent when the MTC devices are farther from the MeNB.

The data delivery delay for MTC devices is shown on the right. It is clearly shown that the performance deteriorates in case A. On the contrary, through the HeNB, the delivery time is decreased by a factor of 14 percent until the distance of 400 m, while the gain increases up to 30 percent when the distance becomes greater. The improvement is due to the fact that, being at a shorter distance from the HeNB w.r.t. the distance from the MeNB, MTC devices experience better channel qualities, and consequently data transmission can be handled in a more efficient way by exploiting less robust transmission parameters (i.e., modulation and coding schemes).

Radio Access Congestion under Ultra-Dense MTC Deployment

The analyses presented so far demonstrate the suitability of supporting MTC via HeNBs in a scenario with a fixed number of MTC devices. Here, we are interested in showing the scalability of our solution with an increasing number of MTC devices. When a very large and unpredictable number of machines simultaneously attempt to access LTE-A networks, we expect that the use of HeNBs should reduce the radio access congestion. With this aim, we stressed the simulation scenario by:

- Reducing the period for arrivals of both HTC and MTC terminals to 20 s
- Varying the number of MTC devices (from 1000 to 30,000) located at the cell edge (i.e., 800 m from the MeNB)

This scenario also resembles the case of MTC devices located in challenging locations with poor network coverage, such as indoor environments.

Impact on HTC Traffic

The radio access network becomes the bottleneck of an LTE-A system when a very large number of MTC devices transmit within a short time interval. As highlighted in Fig. 4, when HTC and MTC traffic is served via the MeNB only (i.e., case A), the probability of successful data transmission is drastically reduced for both traffic types with an increasing number of MTC devices. Indeed, with numerous MTC devices simultaneously transmitting, the amount of resources available for HTC users quickly becomes scarce. As a consequence, HTC users are not able to successfully complete their own data transmission. It is worth noticing that network congestion is already evident with 3000 MTC devices; with above 5000 MTC devices, the transmission of BE flows is denied in the cell.

When we consider case B, we obtain a totally different behavior. Indeed, the HeNB avoids MTC services to take resources from the HTC traffic, which are always successfully transmitted (i.e,. success probability equal to 1 for HTC and MTC traffic). In addition, also for the case with 30,000 MTC devices, the HeNB is able to handle the MTC traffic, confirming the effectiveness of the proposed solution in extremely dense network scenarios.

Design drivers	• Reduced complexity thanks to the frequency-separated scenario
RA procedure	• Higher success probability for the transmission of the first preamble • Reduced latency (i.e., switching time from idle to connected mode) • Reduced energy consumption
Data transmission	• Reduced latency • Reduced energy consumption • Higher capacity (i.e., number of supported MTC devices) • No negative impact on HTC services
Core network	• High scalability (increase in the number of HeNBs does not cause network overload) • No overload at the MME thanks to the use of HeNB-GW • No overload at the S-GW thanks to the use of HeNB-GW • Paging optimization at the HeNB-GW • Inter-connection with trusted non-3GPP APs
M2M uptake	• Decreased dependence on coverage provisioning in challenging environments • Not losing the ability to use operator's connectivity and service platforms • Significantly quicker time to market because of reduced telco engineering work • Significantly reduced risk due to availability of SLAs in licensed bands

Table 1. A summary of the main benefits of the proposed network architecture for MTC over HeNBs in LTE-A.

Impact on MTC Traffic

In Fig. 5, the MTC performance is shown to underline the meaningful improvements achieved by femtocells. In addition to cases A and B, we further consider a scenario (case C) with *mixed MTC and HTC traffic over femtocells*. In detail, in case C we assume that the HeNB is exploited to serve MTC devices, five VoIP and five BE users (the dependence on the number of voice/data users in the HeNB has been omitted here for space reasons). We first consider the data delivery delay (left side). When MTC devices are attached to the MeNB (i.e., case A), the high network congestion causes a quick increase in the data latency. Indeed, with 3000 devices, the delivery delay is already higher than 2 s and is close to 14 s for the case of 30,000 machines. On the contrary, in cases B and C the data delivery delay is lower than 60 ms until there are 20,000 machines. We can observe that, when considering 30,000 devices, the presence of HTC traffic in case C introduces an MTC delay increase of about 200 ms compared to case B. Focusing on the performance of HTC users in case C (results are not plotted due to the lack of space), both VoIP and BE terminals are always able to accomplish data transmission with some performance degradation. In particular, VoIP users experience an increase in the percentage of packet losses up to 20 percent for the extreme scenario with 30,000 MTC devices, while BE users observe a throughput decrease from 670 kb/s (scenario with 1000 MTC devices) down to 560 kb/s (scenario with 30,000 MTC devices).

We also considered a fundamental parameter for MTC: the energy consumption. We are assuming that the MTC devices transmit 1 packet/min (which is rather a high-load case), and their battery capacity is equal to 1500 mAh. By analyzing the *lifetime* of MTC devices in the right plot of Fig. 5, we observe that the increase in data delivery delay due to network congestion in case A involves a drastic reduction in terms of battery lifetime (i.e., from 492 to 17 days when the number of MTC devices varies from 1000 to 3000). The exploitation of HeNB allows *significantly longer battery life* — 625 days in the case of 20,000 machines for both cases B and C. When the number of MTC devices increases, the data delivery time consequently becomes larger and the battery lifetime is reduced to only tens of days.

Concluding Remarks

Conclusions

We present and quantify the performance of a novel network architecture able to efficiently handle ultra-dense MTC over LTE-A networks by exploiting closed access femtocells. Such an architecture allows meaningful advantages, which are summarized in Table 1. Focusing on the coexistence of HTC and MTC traffic, the proposed solution allows effective separation of traffic and, as a consequence, fulfills the requirements of MTC without affecting the performance experienced by HTC users. Concerning the enhancements on MTC devices, the exploitation of HeNBs is useful in offering several benefits such as **significantly reduced latency and energy consumption** in RA and data procedures, and higher capacity (i.e., number of supported machines).

Finally, it is worth noting that the proposed solution with frequency-separated deployment for MeNB and HeNBs guarantees reduced complexity (and lower cost) for MTC devices. On the core/backhaul network side, the exploitation of the HeNB-GW in the proposed architecture has the main benefits of reducing the congestion/overloading of both the MME and S-GW in the core. In addition, the proposed network offers connectivity with reduced latency among LTE-A and non-3GPP MTC devices via trusted non-3GPP APs.

Open Challenges

Open challenges relevant to the capability of LTE-A to support the extremely dense scenarios expected for MTC remain.

Latency reduction: According to the analyses shown above, the exploitation of HeNBs is a valid solution to reduce the latency of MTC devices. However, we should consider that the RA procedure involves a very large overhead for data transmission, mainly due to two factors:

- The high latency required for RA before effective data transmission
- A large number of control bits transmitted before the transmission of typical small packets

To solve such issues, an enhanced mechanism for RA and data transmission procedures are required to further reduce latency and energy consumption, and to enhance the efficiency.

Energy-aware radio access: The design of novel techniques to improve the energy efficiency and performance of the RA procedure is still considered an open challenge. In this direction, effective solutions are still needed to guarantee energy-aware RA, that is, an ad hoc RA procedure where the residual energy-life of MTC devices is taken into account to guarantee higher RA probability for those devices with limited battery charge.

Focusing on the coexistence of HTC and MTC traffic, the proposed solution allows for effective separation of the traffic and, as a consequence, fulfills the requirements of MTC without affecting the performance experienced by HTC users.

Transmission of alarm messages: A portion of MTC applications are based on the transmission of alarm messages, which should be conveyed to remote servers or actuators with very strict time constraints. The transmission of alarm messages still represents an open issue since effective solutions are required to guarantee reduced (and deterministic) latency for alarm messages.

Overall, however, we address the arising problems of massively dense MTC deployments through the architectural changes proposed and quantified in this article.

Acknowledgment

The research of Massimo Condoluci is supported by the European Union, European Social Fund, and Calabria Regional Government. This article reflects the views only of the authors, and neither the EU nor the Calabria Regional Government can be held responsible for any use that may be made of the information contained herein.

References

[1] 3GPP TS 22.368, "Service Requirements for Machine-Type Communications (MTC)," Rel. 12, Dec. 2013.

[2] D. Astely *et al.*, "LTE Release 12 and Beyond," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013, pp. 154–60.

[3] V. Gonclves and P. Dobbelaere, "Business Scenarios for Machine-to-Machine Mobile Applications," Int'l. Conf. Mobile Business and 9th Global Mobility Roundtable, June 2010, pp. 394–401.

[4] J. Alonso and M. Dohler, "Machine-to-Machine Technologies & Markets - Shift of Industries," Tutorial, IEEE WCNC, 6 Apr. 2014, Istanbul, Turkey.

[5] K. Zheng *et al.*, "Radio Resource Allocation in LTE-Advanced Cellular Networks with M2M Communications," *IEEE Commun. Mag.*, vol. 50, no. 7, July 2012, pp. 184–92.

[6] 3GPP TR 37.869, "Study on Enhancements to Machine-Type Communications (MTC) and Other Mobile Data Applications; Radio Access Network (RAN) Aspects," Rel. 12, Sept. 2013.

[7] L. M. Ericsson, "More than 50 Billion Connected Devices," 2011.

[8] ETSI TS 123 682, "Architecture Enhancements to Facilitate Communications with Packet Data Networks and Applications," Sept. 2013.

[9] J. G. Andrews *et al.*, "Femtocells: Past, Present, and Future," *IEEE JSAC*, vol. 30, no. 3, Apr. 2012, pp. 497–508.

[10] I. Widjaja and H. La Roche, "Sizing X2 Bandwidth for Inter-Connected eNBs," IEEE VTC-Fall, Sept. 2009, pp. 1–5.

[11] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN)" Rel. 12, Dec. 2013.

[12] 3GPP TS 23.402, "Architecture Enhancements for Non-3GPP Accesses," Rel. 12, Mar. 2014.

[13] 3GPP TR 25.814, "Physical Layer Aspect for Evolved Universal Terrestrial Radio Access (UTRA)," Rel. 7, Sept. 2006.

[14] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, Dec. 2013, pp. 4–16.

[15] 3GPP TR 38.868, "RAN Improvements for Machine-Type Communications," Rel. 11, Oct. 2011.

Biographies

Massimo Condoluci (massimo.condoluci@unirc.it) received his B.S. (July 2008) and M.S. (July 2011) degrees in telecommunications engineering from the University Mediterranea of Reggio Calabria, Italy, where he is currently a Ph.D. student in telecommunications engineering. His main research activities focus on radio resource management and multicast services in fourth generation cellular networks.

Mischa Dohler [F] (mischa.dohler@kcl.ac.uk) is a full professor in wireless communications at King's College London, head of the Centre for Telecommunications Research, co-founder and member of the Board of Directors of the smart city pioneer Worldsensing, Distinguished Lecturer of the IEEE, and Editor-in-Chief of *Transactions on Emerging Telecommunications Technologies*. He is a frequent keynote, panel, and tutorial speaker. He has pioneered several research fields, contributed to numerous wireless broadband and IoT/M2M standards, holds a dozen patents, organized and chaired numerous conferences, has more than 200 publications, and has authored several books. He has a citation h-index of 37. He acts as a policy, technology, and entrepreneurship adviser, examples being Richard Branson's Carbon War Room, the House of Lords of the United Kingdom, the EPSRC ICT Strategy Advisory Team, the European Commission, the ISO Smart City working group, and various startups. He is also an entrepreneur, angel investor, passionate pianist, and fluent in six languages. He has talked at TEDx, and has been covered by national and international TV and radio; his contributions have been featured on BBC News and in the *Wall Street Journal*.

Giuseppe Araniti [M] (araniti@unirc.it) is an assistant professor of telecommunications at the University Mediterranea of Reggio Calabria, Italy. From the same university he received his Laurea (2000) and Ph.D. degrees (2004) in electronic engineering. His major area of research includes personal communications systems, enhanced wireless and satellite systems, traffic and radio resource management in 4G mobile radio systems, multicast and broadcast services, and digital video broadcasting-handheld.

Antonella Molinaro [M] (antonella.molinaro@unirc.it) is an associate professor of telecommunications at University Mediterranea of Reggio Calabria. She was with the University of Messina (1998–2001) and University of Calabria (2001–2004) as an assistant professor, and with the Polytechnic of Milano (1997–1998) on a research contract. She worked at Telesoft, Rome (1992–1993) and Siemens, Munich, Germany (1994–1995) on a European fellowship contract. Her current research interests focus on vehicular networking and future Internet architectures.

Kan Zheng [M'03, SM'09] (kzheng@ieee.org) received his B.S., M.S., and Ph.D. degree from Beijing University of Posts & Telecommunications, China, in 1996, 2000, and 2005, respectively, where he is currently an associate professor. He worked as a senior researcher at companies including Siemens and Orange Labs R&D (Beijing), China. His current research interests lie in the field of machine-to-machine (M2M) communication, cooperative communication, and heterogeneous networks.

Distributed Mobility Management for Future 5G Networks: Overview and Analysis of Existing Approaches

Fabio Giust, Luca Cominardi, and Carlos J. Bernardos

Fabio Giust and Carlos J. Bernardos are with University Carlos III of Madrid.

Luca Cominardi is with University Carlos III of Madrid and IMDEA Networks Institute.

The research leading to these results has received funding from the European Community's Seventh Framework Program FP7/2007-2013 under grant agreement 317941 — project iJOIN. The European Union and its agencies are not liable or otherwise responsible for the contents of this document; its content reflects the view of its authors only.

ABSTRACT

The ever-increasing demand of mobile Internet traffic is pushing operators to look for solutions to increase the available bandwidth per user and per unit of area. At the same time, they need to reduce the load in the core network at a reasonable cost in their future 5G deployments. Today's trend points to the deployment of extremely dense networks in order to provide ubiquitous connectivity at high data rates. However, this is hard to couple with the current mobile networks' architecture, which is heavily centralized, posing difficult challenges when coping with the foreseen explosion of mobile data. Additionally, future 5G networks will exhibit disparate types of services, posing different connectivity requirements. Distributed mobility management is emerging as a valid framework to design future mobile network architectures, taking into account the requirements for large traffic in the core and the rise of extremely dense wireless access networks. In this article, we discuss the adoption of a distributed mobility management approach for mobile networks, and analyze the operation of the main existing solutions proposed so far, including a first practical evaluation based on experiments with real Linux-based prototype implementations.

INTRODUCTION

In the recent years, Internet data communications have experienced a paradigm shift from the traditional fixed cable access to the wireless and mobile world. The huge success of powerful handheld devices and the deployment of faster heterogeneous radio access technologies, like IEEE 802.11n and Long Term Evolution (LTE), have led to the familiar concept of being *connected anywhere*, *anytime*. Reports such as [1] show that mobile traffic growth will not decelerate; conversely, it will increase 11-fold from 2013 to the end of 2018.

Mobile operators, together with industry and research communities, are looking at cheap and effective solutions to cope with this tremendous growth. There are two main issues to tackle:

- How to provide enough capacity in the access
- How to handle all the traffic in the transport network

For the first issue, reducing the size of cells is the most feasible approach that can provide a significant bandwidth increase. Regarding the second issue, current architectures for mobile and cellular networks are highly centralized and hierarchical, forcing user traffic to traverse all the network parts up to the core, where key entities are deployed to function as border IP gateways and mobility anchors. Following this approach, the general packet radio service (GPRS) Tunneling Protocol (GTP) [2] and Proxy Mobile IPv6 (PMIPv6) [3] have been adopted as two possible choices to operate the Evolved Packet Core (EPC) of 4G networks. The advantage of the centralized approach resides in its simplicity, because the central anchor can follow user movements by simply rerouting the packets over tunnels created with the access router where the mobile node (MN) is currently connected. However, the mobility anchor represents a single point of failure, poses scalability issues (i.e., it is the cardinal point for the control and data plane for millions of users), and, in general, leads to suboptimal paths between MNs and their communication peers (also known as correspondent nodes, CNs) [4].

Therefore, future 5G mobile networks are expected to be more flexible, relaxing the constraint of binding user traffic to a central core entity and allowing Internet services to be located closer to the users. Extremely dense wireless deployments shall benefit from such features by reducing the congestion in the operator's core infrastructure and providing improved service to users. Another defining characteristic of future 5G networks is that the infrastructure is expected to simultaneously serve very different sets of users and applications. For example, 5G networks are foreseen to share resources to cope with both highly demanding video applications of a few mobile users and low-bit-rate traffic from a large bunch of sensors (the so-called Internet of Things, IoT). Along with these objectives, distributed mobility management (DMM) has recently emerged as a new paradigm to

0163-6804/15/$25.00 © 2015 IEEE

design a flat and flexible mobility architecture, allowing traffic to be broken out locally closer to the edge (i.e., offloading the network core) and exploiting the use of different gateways for traffic with different connectivity and mobility requirements.

In this article, we argue that DMM approaches are suitable candidates for mobility management in future 5G very dense deployments. Then we explore the DMM solution space by focusing on the main three families of solutions currently proposed:

- A protocol derived from a classical IP mobility management approach, PMIPv6
- A mechanism based on software defined Nnetworking (SDN)
- A routing-based solution

We describe in this article the main characteristics of each of these DMM approaches and then conduct a validation and performance assessment of each of them by implementing the three solutions in a real prototype. Finally, we derive some interesting conclusions from the comparison of the obtained results. In this work we focus on the comparison of DMM-only solutions, but readers interested in a centralized vs. distributed study might consider the analysis reported in [5].

Distributed Mobility Management

The deployment of extremely dense radio networks addresses the need to expand the network capacity, offering an increased bandwidth per user per unit of area. The cellular pico and femto cells, in conjunction with the new advances in the IEEE 802.11 family, like the .11n and .11ac amendments, are speeding up the development in this direction.

Within this context, the current mobile architecture's centralized model poses some scalability issues due to traffic and signaling handling. For instance, in the Evolved Packet System (EPS) architecture, traffic generated in the radio access network (RAN) is conveyed by intermediate nodes called serving gateways (S-GWs) to the packet data network gateway (P-GW) by means of tunneling. The P-GW hence aggregates the traffic from several edge networks and acts as a gateway between the operator's network and external IP networks. While the deployment of extremely dense wireless networks tackles the expected traffic growth in the access part, a solution is also necessary for the core. In this sense, a flatter mobile network is best suited, as it would permit traffic to be routed without traversing core links unless necessary. Moreover, future 5G networks will simultaneously serve traffic from multiple devices with disparate requirements, as, for example, the IoT is expected to increase its footprint in the coming years. This fact requires more flexible network architectures capable of coping with multiple flows with different requirements, and dynamically adapting to the current demands.

The Third Generation Partnership Project (3GPP)[1] has already started developing solutions for the EPS to avoid tying IP connections to core gateways, like the Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO) techniques [6] and LIPA Mobility and SIPTO at the Local Network (LIMONET) [7]. Consequently, IP networks appear closer to user terminals, since the complex operator's backhaul and core infrastructure might be bypassed. Very dense wireless networks could take advantage of this scenario, as they can be deployed in campuses, malls, transportation systems, and so on, which can benefit from having a locally available connection to Internet services (called a local breakout point) so that traffic generated locally is not forced to pass through the core network. In addition, users should join and leave any of these networks without experiencing any service interruptions, enjoying transparent mobility support for those applications that require so.

The DMM paradigm embraces the concepts expressed above, aiming at designing a flat mobile architecture that enables enhanced access to IP services and built-in support for mobility and heterogeneous radio access technologies [4, 8, 9]. The DMM framework envisions an all-IP infrastructure where users' data flows are routed through the optimal path, exploiting multiple anchor points and deployment of IP services closer to the users. Note that this framework envisions supporting mobility across heterogeneous networks, without requiring complex dedicated support from MNs. In addition, a wise assignment of IP addresses to MNs according to the available services for each user provides a mobile operator with the flexibility to handle users' data traffic according to an extended set of policies, such as whether IP flows should be anchored locally (e.g., for short-term sessions) or to a centralized node (for long sessions). This feature — which is very attractive for future 5G deployments, as previously discussed — is known as *prefix coloring* [10], and consists of attributing some meta-data to the assigned prefixes so that they can be used to access particular services, differentiated for the geographical location or by other means depending on the operator's policy.

Description of the DMM Solutions

There have been several DMM approaches proposed so far, spanning from extensions of current standardized protocols to clean-slate solutions. In this section, we describe the operation of the three main families of DMM approaches. These are the most important types of solutions, approaching the elimination of a single mobility anchor from disparate perspectives. Since there are more than one possible solution fitting each of the families, we have selected one per category, and we argue that the obtained conclusions also apply to any other solution from the same family.

The first family of solutions is based on modifications of classical IP mobility protocols, in particular of the well known PMIPv6; thus, in the following we refer to the solution belonging to this family as the *PMIPv6-based DMM solution*; the second category follows an SDN paradigm, so we call the protocol from this cate-

Since there is more than one possible solution fitting each of the families, we have selected one per category, and we argue that the obtained conclusions also apply to any other solution from the same family.

[1] http://www.3gpp.org/

In our PMIPv6-based DMM solution, the MAG role is replaced by the DMM-Gateway. A DMM-GW evolves from a MAG as it is provided with links to the Internet that do not imply paths traversing the LMA. Hence, the DMM-GW acts as a plain access router to forward packets to and from the Internet.

gory an *SDN-based DMM solution*; and the third design leverages on IP routing protocols, hence the name *routing-based DMM solution* for the mechanism within this family. The three groups are made of network-based mobility management protocols, so no mobility client is required on the terminal. However, the first and third groups make extensive use of existing Internet Engineering Task Force (IETF)[2] standards, whereas the second is a clean-slate approach. All of them share the concept that the access router not only provides connectivity to the MNs (by being their default gateway), but are also enhanced with some specific DMM features. For this reason, throughout the text we refer to a DMM-enabled access router as a DMM-Gateway (DMM-GW).

PMIPv6-Based DMM Solution

In the next paragraphs we present a simplified description of the PMIPv6-based DMM solution described in more detail in [11], where the full protocol details can be found. Since this solution inherits many of its features from Proxy Mobile IPv6 (PMIPv6), we briefly describe this latter first. Proxy Mobile IPv6 is a centralized mobility management protocol where a core entity called the local mobility anchor (LMA) establishes bidirectional tunnels with mobility access gateways (MAGs) located in the access networks. Users' upstream data packets are collected by the corresponding MAG and sent through the tunnel to the LMA, which in turns forwards them to the Internet. Similarly, downstream packets are first received by the LMA, which then dispatches them through the tunnel terminating at the MAG to which the MN is currently attached. By using dedicated signaling messages, called Proxy Binding Update (PBU) and Proxy Binding Acknowledgement (PBA), between the MAG and the LMA, the PMIPv6 protocol coordinates the status of the network, letting the LMA know at which MAG an MN is connected to properly route its traffic. Indeed, since the LMA is traversed by users' data flows, it is straightforward for it to redirect the packets to the appropriate tunnel upon handover, based on the indications received from the MAGs. However, in this way, the data path may end up being suboptimal, and the LMA must be provisioned with high-speed and redundant links to the MAGs in order to convey the traffic for all the subscribers.

In our PMIPv6-based DMM solution, the MAG role is replaced by the DMM gateway. A DMM-GW evolves from a MAG as it is provided with links to the Internet that do not imply paths traversing the LMA. Hence, the DMM-GW acts as a plain access router (i.e., no tunneling) to forward packets to and from the Internet. Also, a DMM-GW features mobility anchoring functions, being able to forward without disruption the IP flows that an MN started while attached to it before moving to a new DMM-GW afterward. Moreover, PMIPv6's LMA is reduced to a control plane only entity, referred to as the control mobility database (CMD). The CMD stores, for every MN, all the prefixes advertised to the MN, which DMM-GW advertised each prefix, and to which DMM-GW the MN is currently connected. In addition, by means of extended PBU/PBA signaling, the CMD sends instructions to recover the MN's ongoing IP flows after a handover. As a result, this architecture's scalability is improved with respect to PMIPv6, as DMM-GWs are able of locally breaking out some traffic, thus avoiding the need to traverse the network core. This allows the overprovisioning typically performed when designing the aggregation links from the access to the network core to be reduced.

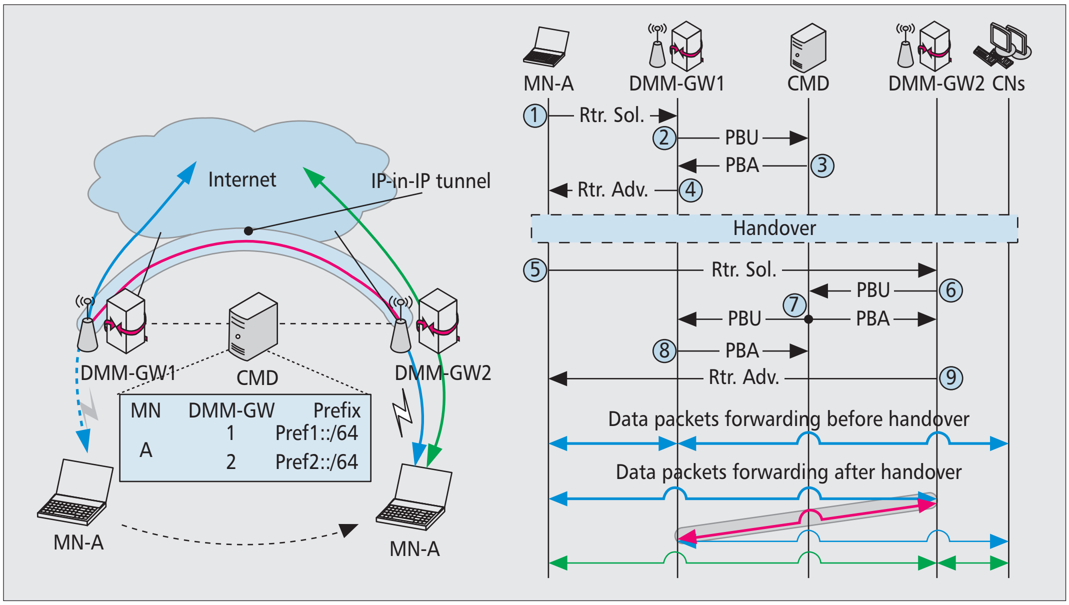
More details and the operations of the PMIPv6-based DMM solution are shown on the right side of Fig. 1, represented by circled numbers. A DMM-GW detects the MN attachment typically after receiving a Router Solicitation message [12] from the MN, or by means of a dedicated link detection mechanism, ①. Next, the DMM-GW notifies the CMD about the MN attachment by means of extended PBU/PBA signaling, ②③, which also contains the IPv6 prefix the DMM-GW is allocating for the MN. Since this is a fresh registration, the CMD creates a new entry for the MN, storing a pointer to the MN's current location (i.e., the DMM-GW that generated the signaling) and a field for the prefix assigned. The DMM-GW advertises the prefix to the MN in a Router Advertisement (RA) message ④ [12]. After a handover, ⑤, when the CMD receives the PBU from the new DMM-GW, ⑥, the database entry for the MN is updated, associating the MN's location with the new serving DMM-GWs. In addition, using the PBU/PBA signaling the CMD instructs the serving and old DMM-GWs to establish a tunnel between them, ⑦⑧. The tunnel is necessary to redirect ongoing IP flows anchored at the old DMM-GW to and from the new DMM-GW. However, the tunnel carries the packets only for those flows that were started before the MN handed over from the previous DMM-GW, whereas new communications are handled by the new DMM-GW as a plain router (i.e., without using any tunnels). This dynamic flow handling is achieved by assigning a new IPv6 prefix to the MN from each DMM-GW to which it connects. The prefix is announced by the DMM-GWs with an RA, ④⑨, which forces the MN to use the new prefix (advertised by the DMM-GW where the MN is currently connected) for new communications. Therefore, each DMM-GW is responsible for a pool of IPv6 prefixes from which it delegates one to each MN attached to its access links. Thereby, a DMM-GW handles users' packets selectively with or without encapsulation, depending on the IPv6 prefix they carry and where the MN is currently connected. Consequently, an MN configures several IPv6 address, one per each visited DMM-GW, and its flows might be anchored at different DMM-GWs.

According to the DMM terminology, this protocol falls within the *partially distributed* category. Indeed, the data plane is distributed among the DMM-GWs, and the control plane is kept centralized, tied to the role of the CMD.

SDN-Based DMM Solution

Software defined networking is a networking paradigm that separates the control and data forwarding planes. Such separation allows for quicker provisioning and configuration of net-

[2] http://www.ietf.org/



MN	DMM-GW	Prefix
A	1	Pref1::/64
	2	Pref2::/64

Figure 1. PMIPv6-based DMM: overview and operations.

With SDN, network administrators can program the behavior of both the traffic and the network in a centralized way, without requiring independent accessing and configuring each of the network's hardware devices.

work connections. With SDN, network administrators can program the behavior of both the traffic and the network in a centralized way, without requiring independent accessing and configuring each of the network's hardware devices. This approach decouples the system that makes decisions about where traffic is sent (i.e., control plane) from the underlying system that forwards traffic to the selected destination (i.e., data plane). Among other advantages, this simplifies networking as well as the deployment of new protocols and applications. In addition, by enabling programmability on the traffic and devices, an SDN network might be much more flexible and efficient than a traditional one.

In SDN environments, the network controller is the most important entity and is responsible for configuring the nodes in the network via a common application programming interface (API), the *Southbound API*. OpenFlow[3] is one API that can be used by an external software application to program the forwarding plane of network devices. The operations of the SDN-based DMM solution are shown on the right side of Fig. 2. In our solution, a core entity, called the network controller (NC), configures the forwarding rules on access routers (the DMM-GWs) using the OpenFlow 1.3 API. The DMM-GWs play the role of anchors. Upon the attachment of an MN to an access point, ①, the DMM-GW informs the NC, ②⑦, which assigns a network prefix to the MN, ④⑤⑨⑩. The network prefix is guaranteed to be unique by using a binding cache where the controller stores, similar to the PMIPv6-based solution, information about the MNs connected to the network. After attachment detection, the NC configures the OpenFlow rules in each DMM-GW visited by the MN, ③⑧.

Mobility is achieved by combining translation and forwarding rules on DMM-GWs. When a packet of an anchored flow reaches a visited DMM-GW, the anchor first rewrites the IP destination address with the last known MN's IP address and then redirects the traffic in the new MN's location. When the traffic reaches the last visited DMM-GW, the DMM-GW performs a reverse IP address translation first, restoring the old IP destination address, and then forwards the traffic to the MN. Note that this solution, unlike the PMIPv6-based one, does not involve any IP tunnels.

This solution, like the PMIPv6-based DMM, is *partially distributed*. While the data plane is distributed, the traffic does not pass through any centralized gateways, and the control plane is centralized at the network controller.

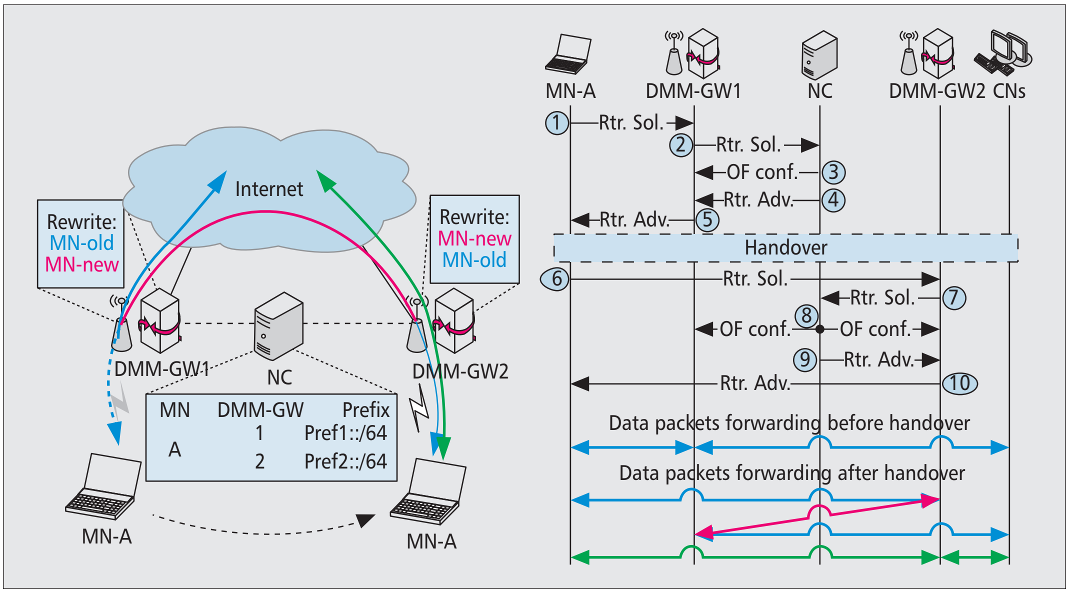
Routing-Based DMM Solution

The basic concept of this type of solutions is to remove any anchor from the architecture, letting all the network nodes re-establish a new routing map when terminals move by means of IP routing protocols. For the purposes of this analysis, we take the solution proposed in [13], which builds on top of the Border Gateway Protocol (BGP) [14] and the Domain Name System (DNS). This is achieved by enabling BGP on the access routers (the DMM-GWs) so that they propagate upward to their BGP peer routers the changes in the access links.

The operations of the routing-based DMM solution are shown in the right side of Fig. 3. Upon an MN attachment to a DMM-GW's access link, ①⑤, the access router learns the MN's DNS name after authentication, ②⑥. Next, the DMM-GW retrieves the IP address (and consequently the IPv6 prefix) associated with the MN's DNS record and announces itself as a next hop to reach the MN's prefix, ③⑦. By doing so, the DMM-GW triggers a BGP routing update in the rest of the network, ④⑧. When the BGP procedure converges, the MN is reachable at the new location using a new path within the network, as depicted in Fig. 3.

It is worth noting that this protocol is fully distributed, in the sense that both the data and control planes are not bound to a specific cen-

[3] https://www.opennetworking.org/



MN	DMM-GW	Prefix
A	1	Pref1::/64
	2	Pref2::/64

Figure 2. SDN-based DMM: overview and operations.

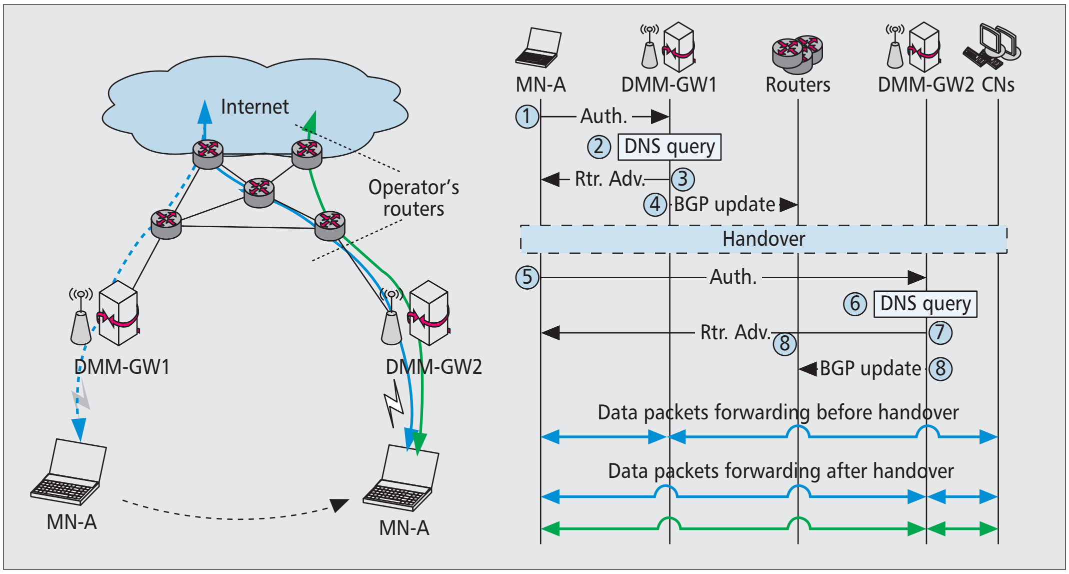


Figure 3. Routing-based DMM: overview and operations.

It should be noted that none of our DMM solutions requires any change on the MN. Indeed, the IP session continuity is provided without any intervention by the MN beyond the Neighbor Discovery operations, which are part of the standard IPv6 stack.

tralized node, but are instead handled by the routers in a distributed way.

Evaluation of the DMM Solutions

After describing how each of the three DMM solutions works, we now report on an evaluation aimed at identifying their advantages and disadvantages, as well as accomplishing an initial performance evaluation.

Table 1 presents a summary of the main characteristics of the DMM analyzed approaches, with a qualitative comparison between them. We next provide some proofs supporting the statements presented in the table.

We have implemented the three solutions described before in order to conduct an experimental performance assessment. The objective of this work is to carry out a proof of concept of these DMM approaches, showing their feasibility with real equipment. Each prototype has been implemented and assessed on the same common platform. The testbed has been realized as a set of three DMM-GWs providing WLAN access using IEEE 802.11b/g cards and a machine acting as CMD, NC, and DNS server. All the nodes run the GNU/Linux operating system. For the sake of clarity, employing 802.11 as the access technology does not affect the functional behavior of the prototypes because, as previously described, the considered DMM approaches are IP-based and layer 2 agnostic. Therefore, the use of other link layer protocols does not have any significant effect on the results.

The PMIPv6-based DMM prototype employs the Mobility Anchors Distribution for PMIPv6 (MAD-PMIPv6)[4] implementation, which runs in the DMM-GWs and the CMD. The code is writ-

[4] http://www.odmm.net/mad-pmipv6

	PMIPv6-based DMM	SDN-based DMM	Routing-based DMM
Type of DMM	Partially distributed (central mobility database)	Partially distributed (SDN controller)	Fully distributed
MN's multiple IP addresses	Mandatory	Mandatory	Supported
Mobility anchors	Multiple (depends on IP flows generation)	Multiple (depends on IP flows generation)	None
IPv6 in IPv6 tunneling	Yes	No	No
Route optimization	No support for anchored IP flows	No support for anchored IP flows	Yes for all IP flows
Handover latency	Low	Low	High
Signaling overhead	Low (depends on no. of active anchors)	Low (depends on no. of active anchors)	High (depends on no. of routers)

Table 1. Features of the three DMM solutions.

ten in ANSI C and provides all the features described earlier.

The SDN prototype employs Open vSwitch[5] as an OpenFlow implementation on DMM-GWs, and Ryu[6] as an OpenFlow-capable SDN framework on the NC. The SDN-based DMM solution is written in Python on top of Ryu's API and provides all the features described previously.

The BGP prototype extends the testbed with a "core" network formed by five routers. The DMM-GWs and "core" routers run the BGP protocol implemented within the Quagga project.[7] An additional piece of software, written in ANSI C, is deployed on the DMM-GWs. This software detects MN attachments and detachments, retrieves the MNs' names and addresses from the DNS server, and installs the local route. When a change in the routing table is detected, the BGP daemon propagates the information to all the other routers. The DNS service is provided by Bind[8] running on the DNS server.

Experimental Results

All the prototypes exhibit three DMM-GWs, each providing WLAN access via a co-located IEEE 802.11b/g access point (AP). The objective of the experiments is to observe how an MN reacts when a handover occurs, that is, what happens to the data traffic when the MN moves from one AP to the other. Therefore, an additional host is deployed, which role is to act as CN, generating *ping* traffic destined to the MN. *Ping* packets must traverse the prototypes once for the request to be delivered to the MN and another time for the reply sent by the MN to the CN. It should be noted that none of our DMM solutions requires any change on the MN. Indeed, the IP session continuity is provided without any intervention by the MN beyond the neighbor discovery operations, which are part of the standard IPv6 stack. We used a laptop as MN, with an out-of-the-box GNU/Linux system (Debian Wheezy OS) and the built-in WLAN card for the wireless access.

For each implementation, we have measured the time required to:

Perform a layer 2 switch, that is, the latency for the MN to change from one AP to another, measured as the time spent for the IEEE 802.11 operations to dissociate from the old AP and associate with the new one.

Perform a layer 3 handover, that is, the time spent since the dissociation from the old AP to the instant when a Router Advertisement is received by the MN, meaning that the MN's IP configuration is ready.[9]

Recover ping traffic, that is, the interval between the last ping packet received or sent by the MN before the handover and the first ping packet received or sent after the handover.

These measurements have been collected by capturing the traffic at the MN's WLAN interface for more than 200 handovers for each platform.

Table 2 showcases the values in milliseconds of the mean value and standard deviation for the three types of handover obtained on the different platforms. Figure 4 depicts in more detail the handover distribution for the ping traffic.

As expected, the layer 2 switch and layer 3 handover are low for all the solutions: this is because the operations performed by the network in order to re-assign the IP connectivity to the MN are very quick. Indeed, in all the schemes, upon detecting the MN attachment, the DMM-GW queries a database in order to retrieve the parameters for the MN's IP configuration. The database is either the CMD, the SDN controller, or the DNS server, respectively, for the PMIPv6-based, SDN-based, and routing-based solutions.

The main difference resides in the time required to recover ongoing data flows, as both the PMIPv6-based and SDN-based solutions are almost 100 times faster than the routing-based protocol. The reason is that the PMIPv6-based and SDN-based mechanisms operate in a conceptually similar way. Indeed, at the time the central database is queried (either the CMD or the NC), this latter sends instructions using the corresponding signaling to the DMM-GWs in order to immediately re-establish a routing path

[5] http://openvswitch.org/

[6] http://osrg.github.io/ryu/

[7] http://www.nongnu.org/quagga/

[8] http://www.bind9.net/

[9] For the sake of simplicity, we do not consider the duplicate address detection process that should be performed after configuring an IPv6 address on an interface.

	Handover type					
	Layer 2 switch		Layer 3 handover		Ping recovery	
	Mean (ms)	Std. sev. (ms)	Mean (ms)	Std. sev. (ms)	Mean (ms)	Std. sev. (ms)
PMIPv6-based DMM	14.0	4.3	26.4	6.7	38.2	10.8
SDN-based DMM	14.0	4.3	35.7	6.7	43.2	7.9
Routing-based DMM	14.0	4.3	59.0	5.9	4743.9	777.1

Table 2. Experimental handover latency results.

for ongoing data flows. In the case of the PMIPv6 solution, the new routing path is achieved with an IPv6-in-IPv6 tunnel between the old anchor and the new one, while in the SDN case new switching rules are installed in the old and new anchors. Hence, both the CMD and NC have an active role in the control plane. On the contrary, the routing-based system does not delegate any control role to the DNS server. Thus, when the access router receives the MN's IP address from the DNS server, the router installs a route for the MN's prefix and sends to the other routers a BGP update notifying itself as next-hop for the announced prefix. Therefore, in order to recover the data flow, all the routers involved in the old data path and those involved in the new target data path must be updated with the correct routing entry, leading to a few seconds latency to let the routing protocol converge. The impact that this might have when the solution runs on large domains is not negligible.

Final Remarks

After giving some figures on the handover latency produced by each solution, it is worth a brief analysis on the protocols' overhead by observing how the signaling messages proliferate during a location update.

In the PMIPv6-based solution, the CMD interacts with the new DMM-GW, and with each of the old DMM-GWs that is anchoring IP flows before the handover. Therefore, the signaling load varies with the traffic and mobility dynamics generated by the MN. In detail, the overhead introduced by the updated location grows with the number of IPv6 prefixes in use by the MN (i.e., the "active" prefixes) because each of them requires a signaling session with the corresponding DMM-GW that assigned the prefix. If the MN's mobility is low, or the IP sessions generated by the MN are short, the number of simultaneous active prefixes is low too, producing little signaling overhead. On the contrary, if the MN is visiting many access networks per unit of time, while keeping several long-lived applications that cannot survive an IP address change, the number of active prefixes is large and so the overhead. However, even if the number of active anchors is large, the latency introduced to recover a communication is impacted only by the distance of the furthest anchor.

The same reasoning applies to the SDN-based approach, leading to the same considerations as for the PMIPv6-based solution.

On the contrary, the number of messages sent by the routing-based DMM solution is determined by the size of the operator's network. Indeed, the DMM-GW needs to notify all of its BGP peers so that the amount of signaling messages sent is almost constant and determined by the number of BGP peers. In a large network, the number of BGP update messages increases dramatically unless adopting some expedients like BGP route reflectors. The time required to re-establish the data communication is affected by the number of BGP routers present in the new data path: as soon as this set of routers converges to the new routing state, the data traffic is recovered. With respect to the service differentiation a future mobile network should offer to the user, we note that the DMM solution proposed in this article may enforce this feature, exploiting a smart IPv6 prefixes assignment by the DMM-GWs. We have observed in the PMIPv6-based and SDN-based solutions that each DMM-GW visited by the MN assigns an IPv6 prefix to the MN used to configure an address at which Internet services can be accessed in a general way. DMM-GWs can assign additional prefixes to the MNs, specifically designated to access some services locally available at that DMM-GW, or addressing some other operator's policies. The routing-based solution is not excluded by this feature, as a DMM-GW can assign a specific prefix, in addition to the main one, to produce service differentiation. Sophisticated use of this technique can lead to a dynamic anchor assignment to the MN's IP flows. For instance, according to the operator's policies, an MN flow can be forced to use a determined prefix for specific IP flows so that the anchoring model (centralized or distributed) can be selected by the operator.

Conclusions

In this article we have focused on DMM as a suitable candidate framework for mobility management in future 5G networks. We have analyzed the DMM solution space by describing the three main solution families for distributing mobility management on a flat architecture for mobile networks.

These three solutions follow different approaches. The first is indeed an extension of a standard mobility protocol for the Evolved Packet System, called Proxy Mobile IPv6. The original protocol has been modified and extended to accommodate a new set of operations, so the outcome is distributed in nature. The second solution operates in a similar way to the previous one, but follows a software defined networking approach. The last mechanism employs the BGP routing protocol to perform the mobility functions required to deliver the packets to and from moving users.

These three types of proposals have been evaluated using real field experiments on Linux-based prototypes. Our findings confirm the intuition that the first two solutions react faster to the changes in the network, but they require dedicated signaling and specialized entities to perform the needed operations. The third mechanism relies on a well established routing protocol, inheriting the issues related to high convergence latency and signaling overhead when used on large network domains.

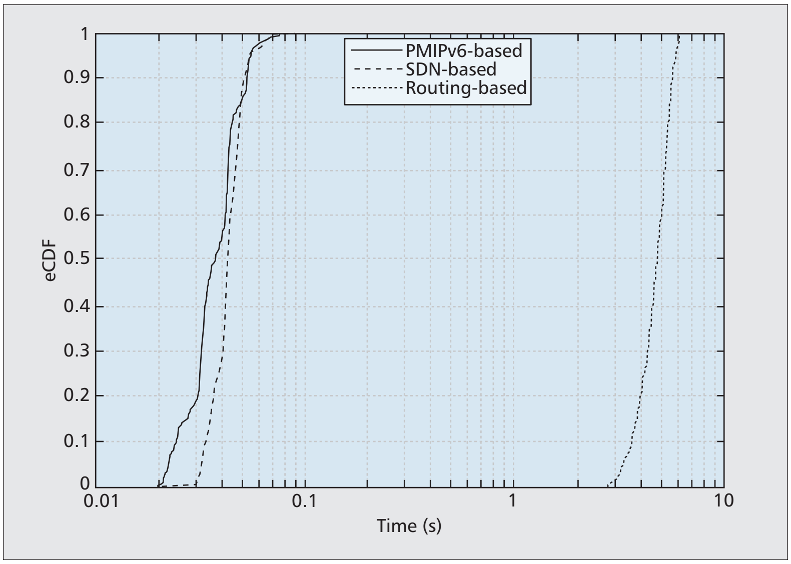


Figure 4. Empirical CDF of the handover measurements for ping traffic.

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018," White Paper, Feb. 2014.

[2] 3GPP TS 29.274, "Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control plane (GTPv2-C)," Sept. 2011.

[3] S. Gundavelli *et al.*, "Proxy Mobile IPv6," IETF RFC 5213, Aug. 2008.

[4] H. Chan *et al.*, "Requirements for Distributed Mobility Management," IETF RFC 7333, Apr. 2014.

[5] F. Giust *et al.*, "Analytic Evaluation and Experimental Validation of a Network-based IPv6 Distributed Mobility Management Solution," *IEEE Trans. Mobile Computing*, vol. 13, no. 11, Nov. 2014, pp. 2484–97.

[6] 3GPP TR 23.829, "Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)," Oct. 2011.

[7] 3GPP TR 23.859, "LIPA Mobility and SIPTO at the Local Network," Apr. 2013.

[8] D. Liu *et al.*, "Distributed Mobility Management: Current practices and gap analysis," IETF Draft, draft-ietf-dmm-best-practices-gap-analysis-07, Sept. 2014.

[9] J. C. Zuniga *et al.*, "Distributed Mobility Management: A Standards Landscape," *IEEE Commun. Mag.*, vol. 51, no. 3, Mar. 2013, pp. 80–87.

[10] M. Le Pape, S. Bhandari, and I. Farrer, "IPv6 Prefix Meta-data and Usage," IETF Draft, draft-lepape-6man-prefix-metadata-00, July 2013.

[11] C. J. Bernardos, A. De La Oliva, and F. Giust, "A PMIPv6-Based Solution for Distributed Mobility Management," IETF Draft, draft-bernardos-dmm-pmip-03, Jan. 2014.

[12] T. Narten *et al.*, "Neighbor Discovery for IP version 6 (IPv6)," IETF RFC 4861, Sept. 2007.

[13] P. McCann, "Authentication and Mobility Management in a Flat Architecture," IETF Draft, draft-mccann-dmm-flatarch-00, Mar. 2012.

[14] Y. Rekhter, T. Li, S. Hares, "A Border Gateway Protocol 4 (BGP-4)," IETF RFC 4271, Jan. 2006.

BIOGRAPHIES

FABIO GIUST (fgiust@it.uc3m.es) received his Bachelor's and Master's degrees in telecommunications engineering at the University of Padova, Italy. After an internship at Alcatel-Lucent Bell Labs in France, he undertook a Master's in telematics engineering at University Carlos III of Madrid (UC3M), Spain. Currently he is working at UC3M, where he is also pursuing his Ph.D. His research interests cover IP mobility and wireless mobile networks, on which he has published several papers in international conferences and journals.

LUCA COMINARDI (luca.cominardi@imdea.org) received his Bachelor's and Master's degrees in computer science at the University of Brescia, Italy. He did an internship and undertook a Master's in telematics engineering at UC3M. Currently he is working at IMDEA Networks Institute and pursuing his Ph.D. at UC3M. His main research interests are SDN, NFV, and integration of the wireless medium into the two former.

CARLOS J. BERNARDOS (cjbc@it.uc3m.es) received a telecommunication engineering degree in 2003 and a Ph.D. in telematics in 2006, both from UC3M, where he worked as a research and teaching assistant from 2003 to 2008 and, since then, as an associate professor. His current work focuses on mobility in heterogeneous wireless networks. He has published over 50 scientific papers in international journals and conferences, and he is an active contributor to the IETF. He has served as Guest Editor of *IEEE Network*.

Software-Defined Networking in Cellular Radio Access Networks: Potential and Challenges

Mustafa Y. Arslan, Karthikeyan Sundaresan, and Sampath Rangarajan

The authors are with NEC Laboratories America Inc.

ABSTRACT

Software-defined networking has brought both performance and management benefits to wired networks. It is natural to wonder if SDN principles also apply and can deliver similar benefits to RANs. It is this question that we intend to address in this article. We first highlight that the core SDN principle of decoupling control and data planes already exists in RANs in the form of self-organizing networking solutions, which can optimize RAN performance at coarse timescales. In addition to control/data plane separation, we also elaborate how fine timescale optimizations such as coordinated multi-point transmission are made practical with the help of cloud RANs (C-RANS), which offer the notion of *processing decoupled from transmission* within the data plane. We finally discuss the potential and challenges related to a less explored application of SDN in the RAN, which is programming the fronthaul network in a C-RAN. We argue that this novel notion of software-defined fronthaul, SDF, has the power to orchestrate novel applications and thus should be a key area of focus in RAN optimization.

INTRODUCTION

Software-defined networking (SDN) has been shown to provide numerous benefits ranging from centralized network management and programmability to reduced capital and operational expenses for wired networks. The fundamental principle in SDN is decoupling the control plane (i.e., configuration and management) and data plane (i.e., forwarding) of the network and facilitating interaction between the two using open interfaces such as OpenFlow [1]. The control plane functions are then aggregated (or centralized) in a software-based controller, which maintains an abstract view of the network topology and provides application programming interfaces (APIs) to network operators. Using the controller and the centralized view it provides, operators can instruct the data plane nodes (e.g., routers) to forward traffic in a certain way depending on their objectives. This is in contrast to the legacy mode of operation, where each node implements both the control and data planes, which makes it difficult to flexibly modify the network behavior since each node has to be programmed separately using potentially different proprietary management interfaces.

Inspired by the above mentioned benefits of SDN, we investigate whether its principles are applicable in the cellular radio access network (RAN) context. We answer this in the affirmative and observe that perhaps the most straightforward application of SDN's control/data plane separation principle to the RAN manifests itself in self-organizing networking (SON) solutions [2]. SON algorithms operate at coarse timescales and optimize RAN performance via control plane coordination without affecting the fine timescale scheduling decisions in the wireless data plane. With the current trend toward dense deployments of small cells (i.e., low-power base stations), such RAN optimizations are essential to ensure the success of future cellular network deployments.

While SON can provide significant performance gains, more sophisticated RAN optimizations require data plane cooperation across base stations at fine timescales. Coordinated multipoint (CoMP [3]) transmission is one such optimization that supports multiple base stations to cooperatively send data to a client, essentially eliminating interference and turning it into cooperation gain. One way to realize CoMP's data plane cooperation is centralizing the data plane but *decoupling processing from transmission* within the data plane to make CoMP practically feasible. This is how cloud RANs (C-RANs) support a centralized data plane by aggregating baseband processing units (BBUs) of base stations and decoupling them from the radio frequency (RF) transmission of remote radio heads (RRHs). In a C-RAN, the BBUs are centralized in a data center — hence the term "cloud" — which hosts general-purpose processors (GPPs) and digital signal processing (DSP) equipment, whereas the RRHs are deployed in a distributed fashion for wireless coverage.

The communication between the BBU pool and the RRHs is provided by the fronthaul transport network, which is a component unique to the C-RAN setup. We show that by program-

0163-6804/15/$25.00 © 2015 IEEE

ming the fronthaul, and thus changing the mapping between BBUs and RRHs at coarse timescales, mobile network operators can both unlock new capabilities of the RAN in terms of traffic optimization as well as obtain energy savings in the BBU pool by powering down BBUs in a dynamic manner. We call this software-defined version of the fronthaul *SDF*, which interestingly resembles the classic notion of SDN in wired networks, and highlight the potential and challenges associated with it.

In summary, we believe that a software-defined RAN should embody the following principles:

- Decoupling the wireless control and data plane to enable coarse timescale SON optimizations that dynamically respond to the changing traffic load and interference levels in the network
- In addition to control/data plane separation, decoupling baseband processing from RF transmission in the data plane through a C-RAN deployment for practical realization of fine timescale interference coordination schemes such as CoMP
- Advancing the C-RAN architecture with an SDF that can be dynamically programmed to enable novel applications

We would like to note that network function virtualization (NFV) is another paradigm, synergistic with SDN, which can be applicable in the RAN context (e.g., base station scheduling virtualization [4–6] or virtualization of the C-RAN BBU cloud). However, our focus in this article is limited to SDN principles and their applications to the RAN.

SON: Control/Data Plane Separation in the RAN

In this section, we first familiarize the reader with SDN and SON by providing a short introduction, and then discuss how core SDN principles can be extended to the RAN within the SON architecture.

A Primer on Control and Data Plane Separation in SDN

In wired networks, while the control plane is responsible for implementing the routing protocols that define the network behavior, the data plane implements the transport (i.e., forwarding) of data across the network. Traditional networks are built with a distributed template where the routers implement both the control plane and data plane functions. Each router is manufactured by a different vendor and thus requires proprietary configuration interfaces. This makes it highly inefficient, if not impossible, for network administrators to configure existing protocols or introduce new innovative network services.

As an alternative to distributed operation where the control and data planes are coupled, SDN proposes an architectural split where the control plane functionality is aggregated (i.e., centralized) in a controller and hence is decoupled from the data plane, which continues to operate in a distributed manner in every network element. The control and data planes communicate using protocols such as OpenFlow [1], which essentially opens up the data plane to be programmed, in software, through open API specifications. Since the data plane is simplified, it can be implemented on general-purpose hardware and controlled with standard protocols, which removes the dependence on proprietary and expensive platforms.

In SDN, programming the network usually involves pushing new forwarding rules down to the switches, which then match these rules based on the features associated with incoming packets. Thus, the data plane only deals with fast rule lookups and executes forwarding at fine timescales, whereas new rules can be pushed into longer timescales due to the latency involved in communicating with the controller.

Each router is manufactured by a different vendor and thus requires proprietary configuration interfaces. This makes it highly inefficient, if not impossible, for network administrators to configure existing protocols or introduce new innovative network services.

A Primer on SON and Its Role in the RAN

SON is an umbrella term describing automated RAN functions, which can be grouped in three categories:

1 **Self-configuration** includes configuration of certain quasi-static base station parameters such as cell ID, neighbor lists, handover thresholds, and management tasks such as software/firmware updates.

2 **Self-optimization** includes dynamic optimization of RAN performance through functions such as interference coordination and load balancing across base stations.

3 **Self-healing** includes failure recovery mechanisms such as adjusting coverage of neighbor cells to serve users of a recently failed base station.

The above mentioned SON functions are especially critical for sustained performance in next-generation small cell deployments. Unlike macrocells, small cells are low-power small-form-factor base stations designed to provide additional coverage and capacity in challenging environments where macrocells do not suffice such as indoors (residential and enterprise buildings) and highly populated urban areas. For increasing the capacity of cellular networks, densely deployed small cells are considered to be of paramount importance. However, due to the shared nature of the wireless spectrum, such deployments bring a number of challenges with them. These include:

- Persistent wireless interference caused by/to the macrocells and other small cells using the same frequency band
- Increased number of handovers for mobile clients since cell ranges are smaller compared to the macrocells
- The need for load balancing between small cells and macrocells for efficient system utilization

Without SON, these functions are extremely difficult and operationally expensive to perform manually.

The Relation Between SDN Principles and SON

In centralized SON implementations, a central controller collects reports about the base stations and their users (using either a proprietary interface or open APIs), configures coarse timescale parameters (e.g., resource assignment to avoid interference), and finally informs the base stations about the configuration using the same reporting interface. Fine timescale operations that happen within a few milliseconds, such as user scheduling in the frame, are executed by the base stations themselves since they cannot be realized in the same report-configure-inform cycle due to latency limitations posed by the interfaces to/from the controller.

In this sense, centralized SON can be seen as a direct extension of SDN's control/data plane separation principle to the RAN (as depicted in Fig. 1). Such solutions are especially important for radio resource management (RRM) problems in small cells. For example, the interference coordination logic (control plane) can be aggregated in the SON controller, which then orchestrates the parameters of base stations (data plane) to help them transmit data to their users without interference. In what follows, we briefly review some centralized SON solutions for interference coordination based on the methodology they use to manage interference in small cells.

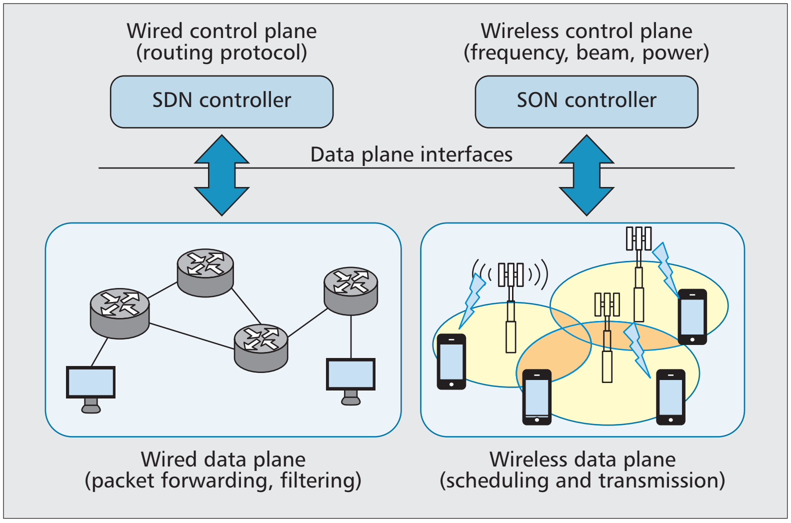


Figure 1. The analogy between SDN and SON.

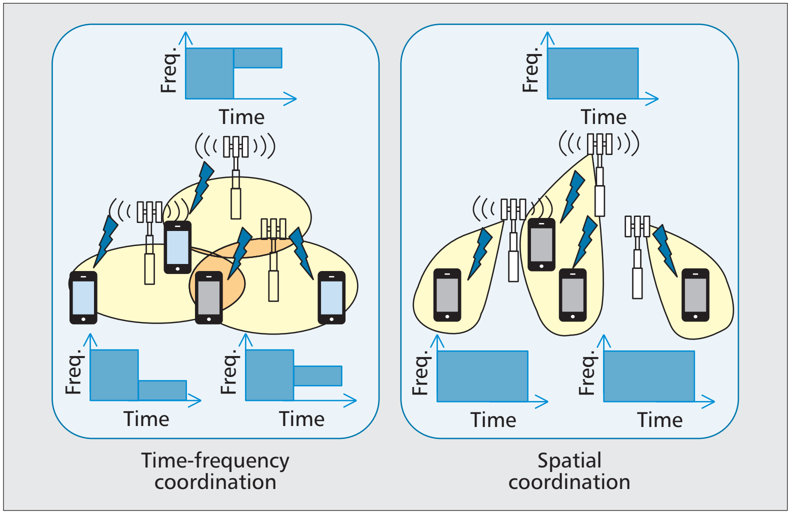


Figure 2. Time-frequency coordination and spatial domain coordination in centralized SON.

Coordination in the Time and Frequency Domains — Cellular technologies such as WiMAX and Long Term Evolution (LTE) have a frame structure where user data transmissions are spread across a 2D grid of time and frequency resources. Thus, interference can be coordinated by either assigning orthogonal frequencies to base stations or letting them time-share the spectrum [7]. For cases where the interference effects are not as severe, one can also exploit resource reuse to assign all of the frequency band to base stations. Reference [8] proposes a hybrid solution by allowing spectrum reuse and adapts resource allocation based on the level of interference and user traffic patterns (Fig. 2).

Coordination in the Spatial Domain — When small cells are equipped with antenna arrays, transmissions can be focused in a particular direction — a technique known as beamforming — by appropriately weighting the signals from each antenna in the array. In addition to improving the signal-to-noise ratio (SNR) at the target receivers, beamforming also reduces interference projected in other directions and thus can be leveraged for interference coordination [9]. Since transmissions are separated in the spatial domain, the base stations can utilize the full capacity of the channel without having to sacrifice time and frequency resources (Fig. 2).

Power Control — Another technique for interference coordination is power control, where each base station adjusts its transmit power to reduce the interference it causes to the users in the neighboring base stations. A central controller can compute the optimal power level for each base station to reduce interference, but it should also account for the performance degradation for the clients served by a base station at reduced power and maintain an acceptable performance level for such users as well [7].

For practical realization in a centralized architecture, interference management should not depend on per-frame scheduling decisions executed by each base station. For this reason, [8, 9] aggregate the interference for all the clients in a cell and perform the optimizations at the cell level (i.e., not on a per-client basis). Since interference is eliminated at the cell level, alternating between different user transmissions in one cell does not impact the level of interference it causes/receives to/from other cells. This decoupling allows such solutions to only compute the optimization in coarse timescales, which is key for practical implementation.

We have shown that centralized SON decouples the control and data planes, and allows for more efficient utilization of radio resources through coordinated control across base stations. However, it is important to note that this design is constrained in two ways:

- The latency in the interface between the SON controller and the base stations limits the optimization to coarse timescales.
- Completely decoupling the data plane gives up the opportunity for *data plane cooperation* to handle interference at fine timescales.

We next discuss a sophisticated mechanism called coordinated multipoint (CoMP) transmission that leverages data plane cooperation and presents unique opportunities for handling interference at much finer timescales than SON.

Practical Data Plane Cooperation with C-RANs

CoMP allows interfering transmitters to share and/or manipulate user data to account for the nature of the interference channel. It transforms the interference into a cooperation gain and can significantly boost the access capacity, especially in dense cellular deployments. There are several schemes under the umbrella of CoMP [3] such as dynamic point selection (DPS), coordinated selection and beamforming (CS/CB), and joint transmission (JT). These schemes allow user data to be simultaneously sent from a set of one or more cooperating base stations. The set of base stations participating in cooperative transmission can be dynamically adapted at the granularity of resource blocks (i.e., time-frequency resources in the frame).

Decoupling Processing from Transmission

It is difficult to realize data plane cooperation techniques such as CoMP on a distributed data plane since the interfaces between base stations pose latency constraints. A practical alternative for implementing CoMP is enabled by C-RANs (depicted in Fig. 3), which centralize the data plane but at the same time provide distributed wireless coverage by *decoupling baseband processing from RF transmission* within the data plane. Unlike conventional RANs, where the baseband units (BBUs) and radio units are situated together, the C-RAN migrates the BBUs to a data center (i.e., the cloud) hosting high-performance general-purpose and DSP processors, while providing high-bandwidth connectivity (called the *fronthaul*) to remote antennas called remote radio heads (RRHs). Aside from the control-/data plane separation in SON, this decoupling of processing from actual transmission within the data plane allows for practical CoMP realization, where the BBUs in the data center can seamlessly share data and cooperate to improve the RAN capacity (several other benefits of C-RAN are detailed in [10]). C-RANs also support two different levels of control for CoMP:

- Coarse timescale control, such as selecting the set of base stations to participate in CoMP, can be implemented in the SON controller.
- Fine timescale controls, such as coordinating the transmissions within the CoMP set, are coupled together with the data plane and thus can be executed in the BBU pool.

In C-RANs, there are various layers at which the BBU-RRH functional split can be implemented. While layer 1 (L1) C-RANs implement baseband processing in the BBUs, L2 C-RANs leave part of the medium access control (MAC) processing to be executed by the RRHs, and in L3 C-RANs, even the network processing is executed by the RRHs. This presents a trade-off in the ability to leverage data plane cooperation and the capacity requirements for the fronthaul. While L1 C-RANs support CoMP, their baseband I/Q signals require the most bandwidth (several gigabits per second) for transport to the RRHs. At the other extreme, L3 C-RANs lose support for CoMP (but SON is supported) but also require the least fronthaul transport capacity.

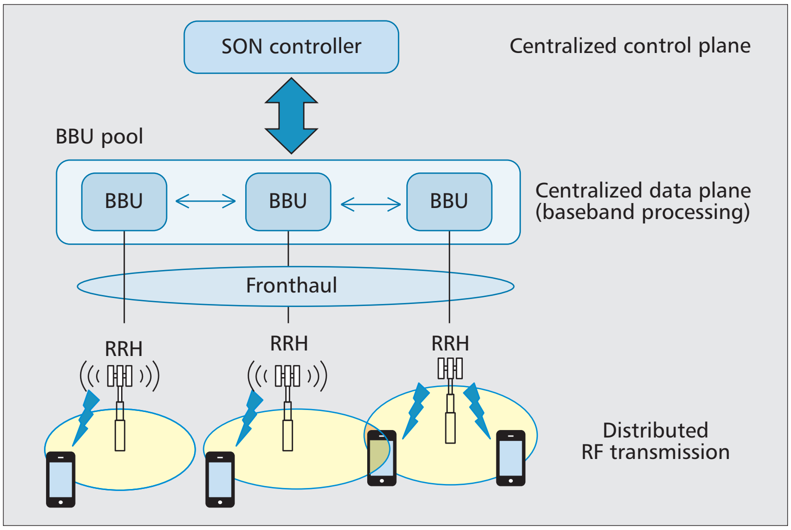


Figure 3. The C-RAN architecture.

One component that has often been overlooked in C-RANs is the fronthaul, which connects the BBUs to the RRHs. We argue that the functionality of the fronthaul should extend beyond merely providing BBU-RRH connectivity. By introducing switching hardware to the fronthaul and flexibly mapping the connections between the BBUs and RRHs, one can enable novel mechanisms that target both traffic optimization in the RAN as well as energy optimization in the BBU pool. We believe that this notion of a programmable fronthaul network, which we call software-defined fronthaul (SDF), closely resembles the classic notion of SDN in wired networks and has not been explored in great detail. We next highlight the potential and challenges related to this novel direction in RAN optimization.

Software-Defined Fronthaul in C-RAN: Potential and Challenges

Although C-RANs decouple the BBUs from the RRHs in terms of physical placement, there is a one-to-one logical mapping between BBUs and RRHs in that one BBU is assigned to generate (receive) a signal (e.g., an LTE frame) to (from) an RRH, although this mapping can change over time. This one-to-one mapping allows for generating a distinct frame for each small cell (deployed in the form of an RRH), which is key for enhancing the network capacity via techniques such as dynamic fractional frequency reuse (FFR [8]) or CoMP [3]. However, this

By introducing switching hardware to the fronthaul and flexibly mapping the connections between the BBUs and RRHs, one can enable novel mechanisms that target both traffic optimization in the RAN as well as energy optimization in the BBU pool.

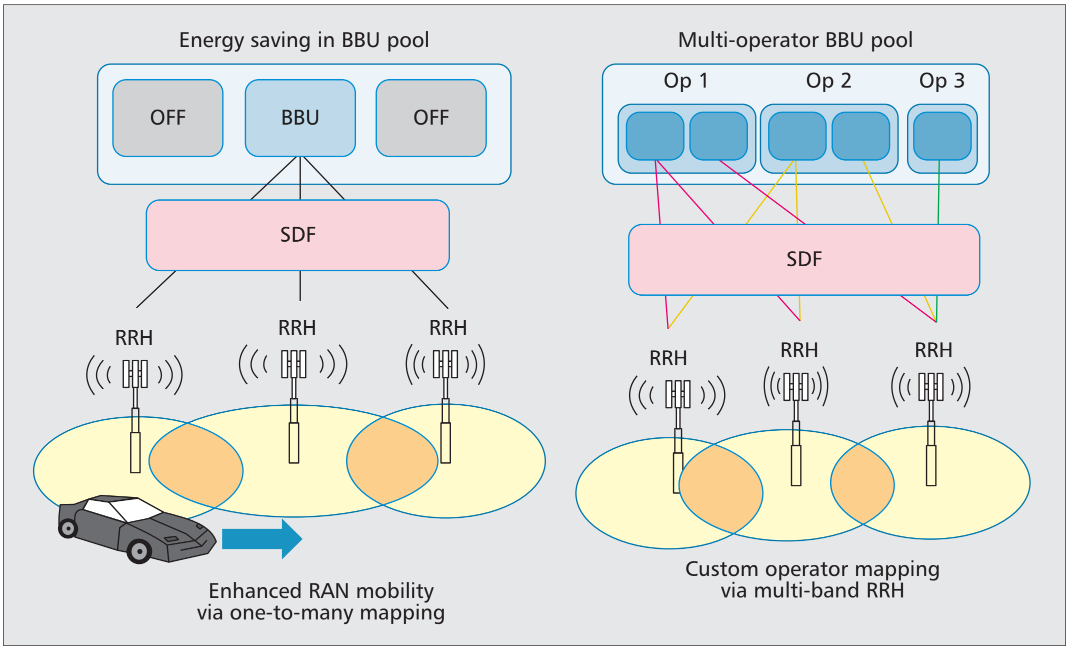

Energy saving in BBU pool
OFF
BBU
OFF
SDF
RRH
RRH
RRH
Enhanced RAN mobility via one-to-many mapping
Multi-operator BBU pool
Op 1
Op 2
Op 3
SDF
RRH
RRH
RRH
Custom operator mapping via multi-band RRH

Figure 4. Illustrations of the potential of SDF.

notion of a fixed one-to-one mapping limits the performance of C-RANs for several reasons, which in turn can be addressed through a *programmable fronthaul network*.

Potential for SDF

Seamless Mobility — RRM techniques such as dynamic FFR primarily apply to static users. Mobile users will have to bear frequent handoffs and the associated performance penalties. In addition, tracking a mobile user's location and channel may be difficult for such techniques. In fact, for mobile clients, a distributed antenna system (DAS) is better suited. In a DAS, the same signal (carrying the user's data) is transmitted simultaneously by multiple small cells to provide coverage benefits (which in turn reduces handoffs) and diversity gain. DAS can be realized by changing the one-to-one to a one-to-many logical mapping in the C-RAN fronthaul, as depicted in Fig. 4.

BBU Energy Consumption — One-to-one mapping requires several BBUs to be active and generating frames, which consumes energy in the BBU pool. However, the enhanced capacity of techniques such as [3, 8] may not be needed in all parts of the network or at all times [11]. When the traffic load is low in a region (e.g., the coverage area of multiple small cell RRHs), a single BBU may suffice to serve the offered load (via a DAS mapping). This in turn reduces the number of BBUs, thereby enabling energy savings in the BBU pool as shown in Fig. 4.

Application Performance — The Evolved Multimedia Broadcast Multicast system (eMBMS) in LTE has provided operators a new revenue stream from video broadcasting and streaming services. eMBMS allows the support of a single-frequency network (SFN), where the same multicast stream is broadcast from all base stations in the SFN, thereby also providing a diversity gain. However, eMBMS requires a back-end server running an eMBMS-specific protocol to synchronize the broadcast content in the SFN. In contrast, C-RANs can provide eMBMS services without eMBMS support from the network. This can be achieved naturally through a DAS mapping, where the same data stream is sent to all cells in the SFN.

Customization for Technologies/Operators — SDF, being standards-agnostic, allows for the coexistence of multiple operators that can share the same fronthaul network, while applying their own fronthaul configurations to cater to their respective user traffic profiles and objectives as shown in Fig. 4. It also allows for different configurations to be applied on different frequency carriers as well as access technologies (WiFi, LTE, etc.) in both single- and multiple-operator scenarios. Such customization is key to fostering innovation in C-RANs.

To summarize, configurations on the fronthaul have a direct impact on wireless transmission techniques (FFR, DAS, CoMP) and hence RAN performance. Also, being the unique component of a C-RAN, orchestrating fronthaul configurations has the power to control both the performance/transmission in the RAN as well as the processing in the BBU pool. Thus, a fronthaul that applies an appropriate mix of one-to-one as well as one-to-many logical mappings between BBUs and RRHs to cater to observed application, traffic load, and user profiles in a software-defined manner, is not only the key to unlocking the true potential of C-RANs but also that of SDN in RANs. Indeed, one can draw an analogy between route computation in wired networks to configuration computation in fronthaul networks from a classic SDN perspective.

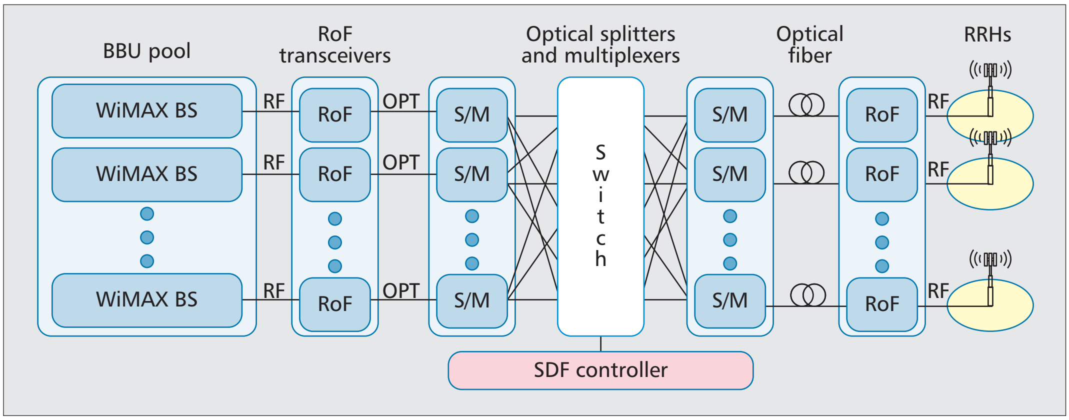


Figure 5. The FluidNet testbed combines radio over fiber with optical switching to realize software-defined fronthaul [12].

Deciding on the switching technology itself is another challenge that must be addressed. Optical switches may incur a longer reconfiguration time than electrical switches but are advantageous in terms of cost, power consumption, and being data rate agnostic.

Realizing SDF

Tantamount to SDN routing in wired networks, SDF requires:

- A controller (resource manager) to compute the appropriate fronthaul configurations for the observed traffic and user profiles
- A switch to effect the configurations in real time

Resource manager: It implements two key functionalities:

- Determining the number of BBU units (based on optimization with network feedback) needed to generate distinct frames and how these frames from BBUs are mapped to specific RRHs
- Assigning compute resources (e.g., CPU cores) to each BBU [13]

Switching element: It realizes the mappings determined by the resource manager. Since some BBU frames are sent to multiple RRHs (as in DAS), while other frames are sent individually to specific cells (as in dynamic FFR), the switching element allows both unicast and multicast. Based on the configuration determined by the resource manager, the switch activates the appropriate set of output ports for an incoming BBU signal depending on the intended set of recipient RRHs. Since a BBU pool may potentially serve tens to hundreds of small cell RRHs, to ensure scalability, the switching fabric may be composed of multiple smaller size switches.

FluidNet [12] is an example of a C-RAN with a reconfigurable SDF that embodies the principles above. Its SDF transports analog RF signals between the BBUs and RRHs through radio-over-fiber technology along with wavelength-division multiplexing (WDM). With WDM, multiple optical carriers can be transported on a single fiber, which is required for the DAS uplink (signals from multiple RRHs going to one BBU) and the multi-operator downlink scenarios (one RRH receiving signals from multiple BBUs). Hence, the switching element in FluidNet is realized in the optical domain, as shown in Fig. 5.[1] In [12], we present real-world experiments and simulations to highlight the benefits of SDF to account for heterogeneity in user behavior (e.g., static vs. mobile users) and heterogeneity in traffic conditions (e.g., low vs. high load).

Challenges for SDF

Building a commercial-grade SDF is a formidable task and faces the following challenges.

Latency: The fronthaul network is responsible for delivering highly delay-sensitive signals between the BBU and the RRHs. For example, each subframe in LTE is 1 ms in duration. Depending on the functional split of the C-RAN deployed, while the fronthaul must support the bandwidth needed to carry the BBU signals, the switch to effect the configurations must do so without affecting the timescale of subframes. Hence, fronthaul switches need to operate with latencies that are an order of magnitude shorter than those of their wired counterparts.

Communication protocol: Since C-RANs are still evolving and have yet to be deployed on a large scale, there is no consensus on open APIs for transport between the BBUs and the RRHs. Furthermore, the nature of the BBU signals that must be handled by the switch will vary depending on the functional split in the C-RAN. There are several common radio protocols (e.g., CPRI) that have traditionally been used to carry BBU signals between the indoor and outdoor units of traditional base stations, and it is possible to repurpose these protocols for fronthaul transport as well. However, integrating such protocols with the switch operations and catering to low latencies is a big challenge.

Electrical vs. optical switching: Deciding on the switching technology itself is another challenge that must be addressed. Optical switches may incur a longer reconfiguration time than electrical switches but are advantageous in terms of cost, power consumption, and being data rate agnostic [14]. These and other trade-offs such as operational cost and reliability need to be carefully evaluated before deciding on a particular technology.

Heterogeneity: Another challenge is heterogeneity, where the interface between the BBUs and RRHs can be a mix of fiber, wireless, and copper links. One needs to integrate

[1] Other realizations of SDF are also possible, where baseband signals are transported in the digital domain.

and efficiently use the bandwidth from the available forms of physical fronthaul to support the logical configurations made by the controller.

Final Thoughts

In this article, we have described the various levels of SDN principles that apply to the RAN:
- Decoupling wireless control and data planes in centralized SON implementations
- Decoupling baseband processing from RF transmission within the data plane in C-RANs to enable practical CoMP
- Equipping the C-RAN with a software-defined fronthaul to enable novel RAN and BBU pool optimizations

Being the least researched component so far, we believe that SDF poses important challenges but also holds great potential for future cellular deployments, and thus should be a key area in RAN optimization.

Being the least researched component so far, we believe that SDF poses important challenges but also holds great potential for future cellular deployments, and thus should be a key area in RAN optimization.

References

[1] ONF, https://www.opennetworking.org/.
[2] 3GPP TS 32.500, "Telecommunication Management; Self-Organizing Networks (SON); Concepts and Requirements."
[3] 3GPP TR 36.819, "Coordinated Multipoint Transmission for LTE Physical Layer Aspects," v. 11.1.0.
[4] R. Kokku *et al.*, "NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks," *IEEE/ACM Trans. Networking*, vol. 20, no. 5, Oct. 2012.
[5] R. Mahindra *et al.*, "Radio Access Network Sharing in Cellular Networks," *IEEE ICNP*, 2013.
[6] 3GPP TR 22.852, "Study on RAN Sharing Enhancements," Release 12.
[7] D. Lopez-Perez *et al.*, "Enhanced Intercell Interference Coordination Challenges in Heterogeneous Networks," *IEEE Wireless Commun.*, vol. 18, no. 3, June 2011.
[8] M. Y. Arslan *et al.*, "A Resource Management System for Interference Mitigation in Enterprise OFDMA Femtocells," *IEEE/ACM Trans. Net.*, vol. 21, no. 5, Oct. 2013.
[9] J. Yoon *et al.*, "ProBeam: A Practical Multicell Beamforming System for OFDMA Small-Cell Networks," *ACM MobiHoc*, 2013.
[10] China Mobile Research Institute, C-RAN: The Road towards Green RAN, http://labs.chinamobile.com/cran/wp-content/uploads/CRAN_white_paper_v2_5_EN.pdf.
[11] U. Paul *et al.*, "Understanding Traffic Dynamics in Cellular Data Networks," *IEEE INFOCOM*, 2011.
[12] K. Sundaresan *et al.*, "FluidNet: A Flexible Cloud-Based Radio Access Network for Small Cells," *ACM MobiCom*, 2013.
[13] S. Bhaumik *et al.*, "CloudIQ: A Framework for Processing Base Stations in a Data Center," *ACM MobiCom*, 2012.
[14] N. Farrington *et al.*, "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," *ACM SIGCOMM*, 2010.

Biographies

Mustafa Y. Arslan (marslan@nec-labs.com) received his B.S. degree from Bilkent University, Turkey, in 2007 and his Ph.D degree from the University of California Riverside in 2012, both in computer science. He has since been a researcher at NEC Laboratories America Inc., Princeton, New Jersey. His research interests include local and wide area wireless networks, with a focus on designing efficient resource allocation techniques to improve performance. He is particularly interested in validation of ideas using proof-of-concept system implementation.

Karthikeyan Sundaresan [SM] (karthiks@nec-labs.com) received his Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta. He is a senior researcher with the Mobile Communications and Networking Research Department, NEC Laboratories America. His research interests span the areas of wireless networks and mobile computing. He currently serves on the Editorial Board of *IEEE Transactions on Mobile Computing*. He was the recipient of Best Paper awards at ACM MobiHoc 2008, IEEE ICNP 2005, and IEEE SECON 2005.

Sampath Rangarajan (sampath@nec-labs.com) received his M.S. degree in electrical and computer engineering and Ph.D. degree in computer science from the University of Texas at Austin in 1987 and 1990, respectively. He heads the Mobile Communications and Networking Research Department, NEC Laboratories America. His research interests span the areas of mobile communications, mobile networks, and distributed systems. He has been on the Editorial Boards of *IEEE Transactions on Computers* and the *Mobile Computing and Communications Review*. He is currently a member of the Editorial Board of *IEEE Transactions on Parallel and Distributed Systems*.

Scalability of Dense Wireless Lighting Control Networks

Conrad Dandelski, Bernd-Ludwig Wenning, Daniel Viramontes Perez, Dirk Pesch, and Jean-Paul M.G. Linnartz

Conrad Dandelski, Bernd-Ludwig Wenning, and Dirk Pesch are with the Cork Institute of Technology.

Daniel Viramontes Perez is with Eindhoven University of Technology.

Jean-Paul M.G. Linnartz is with Philips Research and Eindhoven University of Technology.

ABSTRACT

In modern lighting systems, the introduction of wirelessly controlled LED light sources leads to very dense wireless lighting networks. Current approaches for control message transmission are based on broadcasting messages among many luminaires. However, adequate communication performance — in particular, sufficiently low latency and synchronicity — is difficult to ensure in such networks, especially if the network is part of a wireless building management system and carries not only low-latency broadcast messages but also collects data from sensors. This article describes the challenge of dense wireless lighting control networks. In particular, it discusses the underlying mechanisms and refers to current wireless sensor network solutions in which scalability is relevant.

INTRODUCTION

LEDs are changing the way we illuminate our environment. The efficiency of converting electrical energy into light is not only higher than for conventional lighting, but also the flexibility and controllability leads to a "digitization" of lighting systems. Dynamic control of the intensity, beamwidth, and color point of individual light sources is coming within reach, both technically and economically. Controlling such lighting systems using low-power wireless control networks is attractive due to reduced wiring requirements in such scenarios. This creates a need for low-latency wireless multicast of lighting control messages (one-to-many). Hitherto, building management systems with wireless sensor networks (WSNs) have mostly focused on delivering sensor data to a central node (many-to-one), for example, collection of data about room occupancy, the activities of people, environmental conditions such as temperature, CO_2 levels, and light intensities, where data transmission latency has not been a prime concern for applications such as scheduling of maintenance and janitorial work, improved use of meeting facilities, asset tracking, and saving energy in lighting and heating.

However, with the increasing number of sensors and new applications, self-organization of WSNs, energy-efficient operation of sensor nodes, scalability of networks, and real-time performance for time-critical applications have become critical challenges [1]. In wireless lighting control, for example, a user expects a response within a few hundred milliseconds. In an installation with many light sources, jitter in the communication latency is perceived as a lack of synchronicity between the responses of light sources. This is unacceptable in entertainment applications but can be tolerated to some extent in a parking garage. Moreover, it is necessary that the end-to-end packet loss rate for control message transmission be very low, although the outage probability for a single transmission attempt may be large in any unlicensed band. If an office contains a grid of a few hundred lamps, each controlled via a wireless link, occasional failure of individual lamps to respond needs to be avoided; end-to-end outage probabilities of one or even a tenth of a percent are undesirable, as these are interpreted as "a few lamps are broken."

This article argues the need for more reliable, more scalable dense low-power wireless networks for such applications and the design challenges that need to be overcome.

THE LIGHTING USE CASE

Today we witness many different approaches for wirelessly controlled lighting systems entering the market, but it may be too early to cast a verdict on the most attractive one. Nonetheless, this article summarizes and reviews some of the underlying principles. As the LED wristbands handed out during concerts of the rock band Coldplay prove, multicast/broadcast systems could simply be implemented as a single powerful transmitter that reaches all receivers simultaneously in a single hop. If such powerful transmissions were permitted and frequencies could be reused in all buildings worldwide, it could support arbitrarily high densities, as an arbitrary number of receiving nodes can be added. However, in the more complex systems in building management, where lighting control is just one of many systems that use the wireless communication infrastructure, the combination of broadcasting lighting commands and gathering building management data poses scalability challenges.

0163-6804/15/$25.00 © 2015 IEEE

Figure 1. Typical office setting with multiple light sources.

Lighting networks can be large, as the number of luminaires on a typical office floor is often around 500, but can be as large as 2000 per floor, all part of a single network. However, not all lamps in an office may have to be controlled individually; each lamp may be equipped with its own radio to avoid re-cabling in ceilings of old buildings.

Several network protocols, including ZigBee and 6LoWPAN, can deliver a single broadcast message at sufficiently low latency even if it needs to cross multiple hops. However, in broadcasting by flooding as used in ZigBee, nodes keep repeating this message, sometimes even for many seconds. During this period, newer messages that need to override the first control message may experience a high probability of being lost in collisions. Such effects rule out simple interactive scenarios in which users set the dim level for all lamps and then adjust individual light levels depending on what they see happening.

Besides automated control in response to presence or daylight conditions, it is attractive if an individual user can set their personal light level. The amount of message traffic associated with this is presumably limited to no more than a few messages per hour or per day. In addition to lighting control, building management services demand further sensors to provide data. Their traffic patterns may require periodic data transmissions or be restricted to offloading batches of measured data at hourly intervals without stringent latency requirements. However, they add additional message traffic with which lighting control has to contend.

Where Are We with Dense Low Power Wireless Networks

Channel Access

Effectively, all low-power wireless networks, whether standards-based or proprietary solutions, use either a variant of carrier sense multiple access (CSMA) or time-division multiple access (TDMA) (including the ALOHA type mechanisms) as their medium access control mechanism and essentially suffer from the same problems as the basic mechanisms.

ALOHA, one of the earliest multiple access schemes, grants access to the channel to any node that has a message to transmit. The inherent disregard for other nodes' transmissions results in packet losses and very low throughput, as shown in Fig. 4 .This problem can be mitigated by CSMA (e.g., as used in the IEEE 802.15.4 standard), a common low-power wireless network technology, which operates in the 2.4 GHz industrial, scientific, and medical (ISM) radio band. IEEE 802.15.4 supports star, tree, and mesh topologies, but does not specify how the topologies can be obtained and maintained, so the network formation and routing must be provided by upper layer protocols. However, in star and tree topologies, the network is managed by a personal area network (PAN) coordinator. The beacon-enabled mode, which is restricted to star and tree [2], can achieve better energy efficiency using a superframe structure. The superframe structure is divided into an active (receive/transmit) and inactive (sleep) period. The active period is also divided into a contention access period (CAP), where the devices use slotted CSMA with collision avoidance (CSMA/CA), and a contention free period (CFA), where the devices use seven guaranteed time slots to communicate with the coordinator. The superframe is called a beacon interval (BI), and its length ranges from 15.36 ms to 251.7 s, depending on parameter settings [3].

For example, in a star topology with 21 nodes, a BI of 0.49 s, and interarrival times from 0.05 to 100 s, the end-to-end delay decreases from 2 s to 0.5 s, the packet loss rate decreases from nearly 100 percent to less than 5 percent, while the power consumption for the network increases from 0.02 μAh/byte to 0.15 μAh/byte [2]. In a new approach Feng Chen *et al.* [4] removed the restriction of seven guaranteed time slots, the minimum CAP, and the inactive period; this yields essentially a TDMA-based channel access. With more time slots, the latency decreases to less than 5 ms, and the packet loss rate is below 5 percent at an interarrival time of 0.1 s for 21 nodes in star topology. However, these performance characteristics are too generic and offer little help in regard to the required performance in the lighting control use case.

A TDMA approach is the basis for the IEEE 802.15.4e update released in 2012. The IEEE 802.15.4e specification addresses problems found in industrial automation, office, and home control, such as timeliness, reliability, robustness, scalability, and flexibility, compared to the basic IEEE 802.15.4. With this update the time slotted channel hopping (TSCH) mode was introduced. TSCH uses a superframe structure with n-time slots, where n is the number of time slots for one BI, and up to 16 channels. A typical duration for one time slot is 10 ms (4 ms for transmitting a maximum length frame, 1 ms for acknowledgment, and 5 ms for turnaround). Thus, the latency is predictable. The use of 16 channels allows the use of multiple (sub-) networks to mitigate congestion in individual networks [1]. However, the number of time slots is limited and limits the size of individual networks.

BROADCAST

For several applications, broadcasts must be reliable and efficient in using power and spectrum resources [5]. The simplest form is *classic flooding*. Any node that receives a broadcast message verifies whether the message is new or has already been received. If it is a new message, the node sends the message to its neighbors. Hence, every single node will retransmit each flooded message exactly once. While in sparse networks, classic flooding may fail to reach all nodes, in a dense network it results in *excessive redundancy*, *contention,* and *collisions* [6], known as broadcast storms. Therefore, flooding mechanisms must compromise between minimizing redundancy and maintaining good latency and *reachability* (the ratio of nodes receiving a broadcast message to the total number of nodes that are directly or indirectly reachable from a source node) [5]. In ZigBee, before rebroadcasting, nodes check whether the new message has already been repeated on the channel, while in 6LoWPAN, nodes always instruct their medium access control (MAC) to (try to) retransmit a new message.

Broadcast storms can be mitigated by inhibiting avalanches of retransmissions by estimating redundancy, for example, using probabilistic schemes possibly also guided by message counters or location information [6]. A refinement is to adapt the decision thresholds and parameters for estimation of redundancy to changing network conditions, but retransmissions appear essential to ensure adequate reachability.

Clustering is another approach to mitigate broadcast storms where only specific nodes (e.g., cluster heads) rebroadcast a message, thus reducing the number of nodes involved in transmissions [6]. Here the challenge is to elect the right nodes as cluster heads, which is a nontrivial problem. Generally, in WSNs, broadcast protocols need to adapt how and when they should propagate their information, for which no perfect recipe exists yet. Apart from the MAC layer, message flow can be regulated at the application layer. An example is the use of the Trickle protocol for sharing control settings [7]. It was designed to share messages of software upgrades among a large set of nodes. Each node can decide to perform a retransmission if it does not receive updates from neighboring nodes, or if a node is detected to have outdated information. Thus, Trickle provides an appropriate solution to the requirement that propagation of new control messages must override older ones. What is unclear is its real-time performance in the lighting use case.

ENERGY CONSERVATION

The power consumption of many WSN integrated circuits (ICs) such as for IEEE 802.15.4 is only on the order of tens of milli-amps, but many protocols consume substantial standby power due to idle listening. Duty cycling the receiver is attractive, but typically increases latency. In recent years, much research has targeted energy efficiency in channel access. Many energy-efficient MAC protocols have been proposed, such as low power listening with wake up after transmission (LWT-MAC). If the traffic load increases, this MAC protocol switches from unscheduled to scheduled transmissions. The benefit is higher energy efficiency, but at the cost of delay. If the message interarrival time is below 40 s, the delay increases up to 12 s [8]. Another proposed MAC protocol is scheduled channel polling (SCP-MAC), which is based on low-power listening (LPL). In basic LPL, nodes sleep as much as possible and use an asynchronous wake up mechanism with a long preamble to ensure a transmission is received. SCP-MAC uses a short preamble and synchronized wake-up times to reduce energy consumption. While energy consumption is up to 88 percent lower compared to LPL, the latency for SCP-MAC and LPL reaches 10 s in a multihop network with 9 hops [9]. While LWT-MAC, SCP-MAC, and LPL aim at low energy consumption, their latency prohibits use for real-time lighting applications, leaving the challenge open to create an energy-efficient low-latency low-power MAC protocol.

Figure 2. 300 Philips HUE lamps in a dense ZigBee network shown at the 2014 Light and Building Fair in Frankfurt.

RELATED WORK IN AUTOMOTIVE

Related work on the problem of one-to-many broadcasting in multihop networks with low latency requirements has been investigated for safety message transmission in vehicular ad hoc networks (VANETs). However, the requirements for latency for safety messages in VANETs range from a maximum of 100 ms for “wrong way driver warning” or “collision warning” to a maximum of 1 s for “road condition warning.” Those messages have update intervals with minimums of 1–10 Hz, and the minimum broadcast distances range from 200 to 1000 m depending on the message type. However, the challenge as to how to implement a broadcast function in a distributed sensor network that was originally designed and optimized for multihop forwarding of unicast messages, which involves repetition of messages and resolution of collisions, is relatively unexplored. While some requirements in VANETs are comparable to those for dense WSNs for lighting control, the environment in VANETs differs considerably from WSNs in buildings. Also, the topology in

VANETs is highly dynamic with links appearing and vanishing frequently as cars may move with speeds above 200 km/h. In lighting control, instead, the topology is fixed, and the nodes are mounted in the ceiling, so no path breaks occur, although temporary fades may last a long time. Another difference is in node density. In VANETs the density varies over time from a few equipped cars to tens or hundreds per networked road section, whereas in lighting the node density can be several nodes per square meter and remains constant over very long periods of time.

Commonly Used Standards

A range of low-power wireless communication standards exist as well as a wide range of vendor proprietary solutions that are used in the field of WSNs and wireless control systems. The ZigBee Alliance is an association of companies working together to develop standards and products for cost-effective, reliable, and low-power wireless networking. The ZigBee Alliance provides two ZigBee stacks, ZigBee and ZigBee Pro. The ZigBee stack is designed for smaller networks (hundreds of nodes in a single network) and is compatible with the ZigBee Pro stack, which is designed for larger networks (thousands of nodes). Both ZigBee and ZigBee Pro are working on top of the IEEE 802.15.4 MAC and physical (PHY) layer. ZigBee defines the network layer and provides a framework for the application layer. The network layer organizes the formatting in either star, tree, or mesh topology, and also provides functionalities for multihop routing, route discovery and maintenance, security, and the procedure for joining/leaving a network. There are also many use-case-specific ZigBee standards available, such as ZigBee Building Automation, ZigBee Smart Energy, and ZigBee Light Link. Those standards give additional functionalities for the specific use case [10].

The ZigBee Alliance is an association of companies working together to develop standards and products for cost-effective, reliable, and low-power wireless networking. The ZigBee Alliance provides two ZigBee stacks, ZigBee and ZigBee Pro.

However, ZigBee products are used in many of today's applications worldwide. One example of a large-scale ZigBee network is the Aria MGM City Center Hotel, Las Vegas, Nevada. Over 136,000 ZigBee devices are deployed in the hotel, distributed in over 4000 rooms. In each room different sensors and actuators are installed, such as temperature sensors, light switches, and dimmers. Those sensors and actuators communicate with a control unit in each room and form a local ZigBee network. The control unit of each room is connected via fiber to a centralized server, which controls the whole building. Such infrastructure resembles cellular frequency reuse rather than a dense WSN in the strictest sense, because the 136,000 nodes are separated into 4000 networks, with each network containing around 30 nodes. This approach translates the scalability issue into a cellular reuse problem, which we address later. The number of hops that can be tolerated within one cell is a trade-off between performance and infrastructure cost.

Besides indoor applications for ZigBee and other wireless standards, there are also outdoor applications. One example is a street light control system proposed by Elejoste *et al*. [11]. This streetlight control system is based on the IEEE 802.15.4 standard and uses XBee nodes. In the proposed scenario the light posts are located along a street every 20 m. One streetlight line forms a mesh network (e.g., one street or one block). The lines are connected to a central point that manages all networks. However, this infrastructure does not match the definition of a dense WSN, because of the large distances between the individual nodes.

The Limits to Scalability in Dense Networks

Physical Limitations, Spatial Reuse, and Propagation

Problems arising in dense wireless networks are interference and spectrum efficiency. As dense networks are interference limited, it may be tempting to lower the transmit power to reduce the interference. However, this is often counterproductive: the signal-to-interference ratios (SIRs) do not change if all nodes decrease their transmit power by the same ratio. Thus, the probability of losing messages in packet collisions does not change. However, as the signal-to-noise ratio (SNR) deteriorates by lowering the transmit power, more packets are lost.

A mechanism to increase spectrum efficiency is spatial reuse. One of the early systems that enabled better spatial reuse was the introduction of broadcast frequency modulation (FM) as it tolerates much lower SIR than amplitude modulation (AM). Digital modulation is a further step in making transmissions robust against interference, while digital spread spectrum (CDMA) transmission even allows neighboring cells to use the same channel without a protection area to mitigate interference between co-channel cells. In IEEE 802.15.4 and almost any other standard, relatively narrow channels are defined, and the availability of multiple channels makes the use of different frequencies in neighboring channels attractive. The 21 channels in IEEE 802.15.4 would allow a cluster size and reuse distance far larger than in outdoor cellular networks. Nonetheless, in dense indoor systems this may not provide large SIR levels. Frequency reuse can only work if the path loss falls off sufficiently rapidly with distance. For typical cellular or VANET communication ranges, the received power p typically decreases with $p = d^{-\beta}$ where the path loss exponent β is between 3 and 4 due to out-of-phase ground reflections. It has been widely recognized that a steeper fall-off of the signal strength enhances the spatial capacity of cellular networks. Vice versa, we expect a significant challenge for very dense wireless lighting networks in buildings as the line-of-sight dominated free-space environment, which is typical for dense lighting installations, yields a slow fall-off of only ($\beta = 2$). Hence, such dense networks suffer from interference by many nodes. In large open space offices or parking garages, radio propagation can be guided between the ceiling and the floor, leading to effectively a β of less than 2. This leads to high signal levels at relatively large distances. In moderately large networks this can be an advantage as relatively few hops are needed to reach a destination node. However, in large and very large networks this works adversely: it prevents any effective reuse of the channel in the entire network. However, in lighting control with

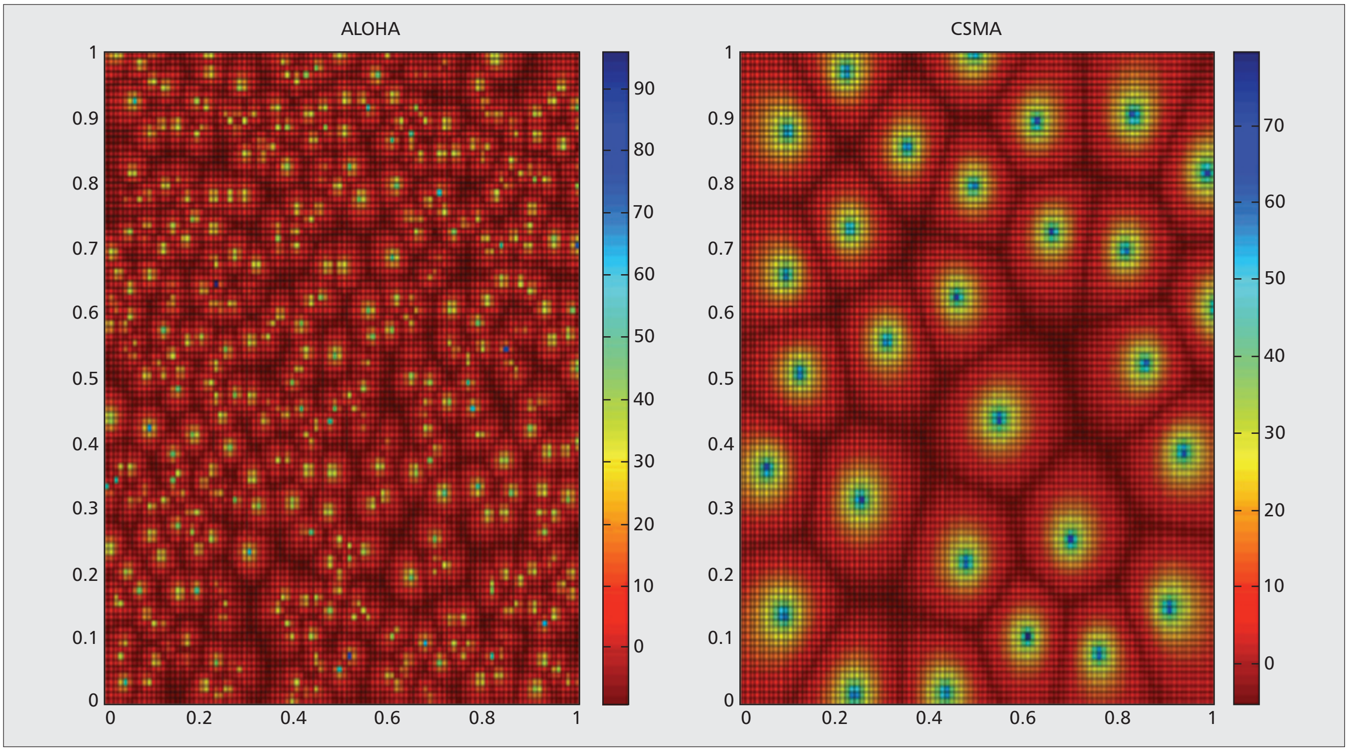


Figure 3. Snapshot of spatial distribution of simultaneous messages.

bursty packetized traffic, the reuse distances can be significantly smaller than in circuit-switched telephony where transmitters are continuously on. In fact, protocol retransmissions for collision resolution within one cluster (cell) can, at least to some extent, also cope with intermittent interference from neighboring co-channel clusters [12]. In practice, it is the too conservative carrier sense setting rather than the excessive collisions that prevents effective spatial reuse. Typically, IEEE 802.15.4 requires a specific choice of the carrier sense threshold, while adaptations to the network lay-out may be appropriate.

In the early days of digital radio communication, spectrum efficiency was often expressed in terms of bits per second per Hertz. However, this figure does not reflect the potential of frequency reuse. A more appropriate measure today is bits per second per Hertz per cubic meter, which highly depends on the minimum SIR z that a radio receiver needs to properly detect a signal among any simultaneously transmitted packets. These signal separation capabilities of the receiver are often simplified into a "vulnerability circle" model by only considering the strongest interferer. Inverting the path loss law for wanted signal and interferer at distances d_0 and d_i, respectively, an interferer that is at a distance $d_i < d_0 z^{-1/\beta}$ causes harmful interference. The converse, the assumption that an interferer outside this range may be neglected, becomes increasingly unrealistic if the network becomes denser. Hence, a capture model based on the signal-to-joint-interference ratio is essential, as confirmed in Fig. 4.

Medium Access Control

For packet data, as in dense wireless building control networks, the learning from cellular radio has two shortcomings: their spectrum is not managed by a licensed operator; second, the transmissions have a bursty nature such that there is no need for a continuous separation between co-channel users.

A revealing thought experiment is a wireless ALOHA network that has access to a channel of bandwidth B such that, according to Nyquist, the packet transmission time T for messages of length m symbols is around $T = m/B$. It needs to accommodate Poisson traffic of λ packets/s, thus the traffic is $G = \lambda T = \lambda\ m/B$. If a collision always corrupts all transmissions involved (no capture), the throughput of successful messages would be $S = G \exp\{-G\} = \lambda\, m/B \exp\{-\lambda\, m/B\}$. If one would split the network into two subnetworks, each using one half of the available bandwidth ($B \to B/2$; thus, $T \to 2T$), and split the traffic over the two subnetworks ($\lambda \to \lambda/2$), the traffic per subnetwork, expressed in number packets per (new) unit of packet time, remains $G = \lambda/2 * 2m/B = \lambda\, m/B$. Hence, creating two subnetworks does not enhance the throughput, but just increases the delay, because $T \to 2T$.

Theoretical Insights for Infinitely Large Network Models

In an ALOHA radio network, collisions occur frequently, but how harmful are these to the performance of broadcasting? In fact, if many nodes retransmit a broadcast message at the same time, a receiver neighboring active transmitters may be able to receive the message anyhow. Theoretical models reveal that the typical communication ranges scale with transmitter density, but that the probability of receiving a message is quite independent of the density of transmitters. Figure 3 shows the SIRs due to simultaneously transmitting nodes according to a

Some regions with bad SINR would be consistently starved from updates. Consequently, the initial advantage in throughput offered by CSMA in comparison to ALOHA might not materialize, as some nodes remain unreached for prolonged times.

Source	# edge routers	# nodes	Avg. # hops	Max # hops	$HN\vartheta$ (kb/s)	Packet success rate	Delay	Sim/exp
Arch Rock	7	220	2.19	5	.6	99.9%		Exp
Arch Rock	4	75	2.67	5	.3	99.9%		Exp
Arch Rock	1	51	2.11	6	2.1	100%		Exp
Arch Rock	5	211	1.20	3	5.1	100%		Exp
RFC 6687	1	2442	5 (?)	12	100 (?)	?	< 1 s	Sim
UBremen	1	20	2 (?)	?	.15	100%		Exp

Table 1. Comparison of network performance for several 6LoWPAN experiments.

spatial Poisson process. Changing the spatial intensity of messages stretches the normalization of spatial dimensions, but cannot change the fraction of red, yellow, and blue regions.

CSMA can enhance the throughput of wireless networks compared to ALOHA by structuring the area as it tolerates transmissions particularly in areas where otherwise the SIR would be poor (red), and inhibiting transmissions near already active nodes (blue), where typically the SIR would be adequate without additional signals. Second, in ALOHA, partial message overlaps can occur for interferers arriving during a time window that has twice the duration of a packet transmission time. CSMA restricts this.

The throughput of an ALOHA network depends on the path loss exponent β. Closed form expressions for the vulnerability circle model [13] and a signal-to-joint-interference model [14, references therein] show that the probability S of receiving a packet is inversely proportional to the ability of the receiver to separate signals, z, raised to the power 2 over β. Theoretically, in a network with an infinitely extended offered traffic that has a spatially uniform intensity of G_0 packet per unit of time and per unit of area, according to Abramson's vulnerability circle model, the throughput equals $S = z^{-2/\beta}$ [13]. Hence, it is independent of the traffic density G_0. The predicted throughput is a factor sinc(2π/β) lower if the accumulation of interference from multiple transmitters, and thus not only from the nearest interferer, is also taken into account.

Interestingly, for Friis free space loss ($\beta = 2$; $p = d^{-2}$), which may be typical in very dense wireless networks where many nodes are within direct line of sight of each other, the throughput decreases to zero ($S = 0$). In such cases the sum of the interference power from all interferers theoretically diverges beyond any finite limit. Figure 4 illustrates that this can have a dramatic effect on very dense networks.

It is unclear whether CSMA always outperforms ALOHA to the extent predicted in Fig. 4. CSMA structures the spatial distribution of traffic, and as the load increases, ultimately a honeycomb-like structure emerges. For sufficiently dense traffic, as soon as one node finishes its transmission, (only) a very close neighbor takes its place. In effect, a *spatial lock-in* could arise [14], preserving the honeycomb structure for a prolonged time. Some regions with bad SINR would be consistently starved from updates. Consequently, the initial advantage in throughput offered by CSMA in comparison to ALOHA might not materialize, as some nodes remain unreached for prolonged times.

Experimental Evidence So Far

There is little hard empirical evidence on what number of nodes a dense wireless network can scale. Results from simulations and testbed experiments have addressed largely many-to-one routing with low traffic intensity. It appears that about 50 nodes per segment controller and up to 5 hops can be supported. Large networks have been demonstrated (e.g., > 400 nodes), but throughput is poor (just a few messages per second) when conditions are near ideal (most nodes in direct contact, thus few hidden nodes).

The total traffic payload in the network can be quantified as the product $N H \vartheta$ with number of nodes N, average hop count H, and per node traffic load ϑ. This figure reflects the claim that an idealized network would put this traffic load on the radio spectrum if it did not incur any overhead. But experimental results report that the total traffic payload $N H \vartheta$ stays well below the link capacity W, which is around 250 kb/s.

Table 1 compares published experiments for 6LoWPAN, typically involving about 50 nodes around one router at the border of the area. Larger networks of up to 2500 nodes have been simulated, but we did not find a confirmation if systems of this size work in practical implementations. Most experiments relied on manually tweaking the routing table generation, and optimization of the routing parameters and packet transmission timing at the application level. Nonetheless, typical average payload rarely exceeds about 2.5 kb/s. Typically we see that $N H \vartheta$ stays below 1 percent of W, so there is a substantial loss in the use of resources.

There are several plausible mechanisms that affect the efficiency of the use of radio resources. First, for very short messages, acknowledgments consume 50 percent of the available resources. However, the random access scheme also affects

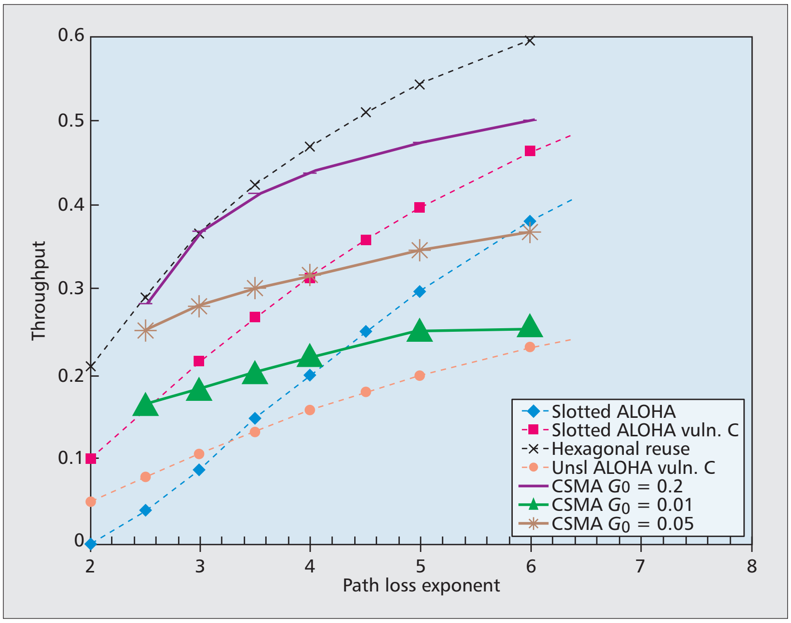


Figure 4. Throughput as a function of path loss exponent β for a large (but not excessive) traffic load per unit of area and $z = 10$. Dense networks typically experience $\beta = 2$.

efficiency. The textbook expression for ALOHA shows a maximum throughput of 1/e (36 percent) for Poisson traffic. This result needs to be interpreted with care, as the expression may not consider wireless channel effects and the maximum of the G exp{–G} curve is not a stable operational point. Typically, queuing systems that operate at their capacity give infinitely large delays to individual users. Wireless systems are no exception. Hence, as a rule of thumb, it is fair to assume that the networks need to be operated at not more than half of the maximum throughput to deliver acceptable delays.

Another substantial factor is the overhead in short messages. The packet header, synchronization preamble, and so on may account for half of the air time.

For broadcasting less experimental evidence is available. In [15] tests in a network with 300 nodes confirmed that messages may well be received correctly despite transmissions that occur simultaneously elsewhere in the network.

Some simulation results can further illustrate how these limitations affect a dense wireless lighting network. In OMNeT++, a simple interactive dimming use case was modeled for a user who turns a dimmer knob and reacts to how s/he sees the light level change. This creates traffic in the form of a sequence of broadcast bursts, transmitted from one node (the dimmer switch). The simulated burst duration of 2 s involves 20 new messages per burst (i.e., one new command every 100 ms). To propagate the message through the network by simple flooding, every node that receives the message retransmits it exactly once. As stated earlier, the different standards are all based on variations of CSMA or TDMA. The MAC layer uses CSMA according to IEEE 802.15.4 and TSCH according to IEEE 802.15.4e. Also, TSCH uses a flooding mechanism, but with the difference that the transmitting node only sends the message to its direct neighbor. Figure 5 reports various network densities ranging from 20 up to 400 nodes, distributed in an office environment with dimensions of 80 m × 40 m.

For any simulated transmit power, the latency increases if the network gets denser, and it also decreases with transmit power. For 0 dBm to –10 dBm the average latency is below 200 ms for all densities, which would fulfil the requirement for wireless lighting control. Reducing the transmit power to –25 dBm leads to excessive latency (above 200 ms) for a density above 200 nodes, as shown in Fig. 5a. Although a reduction of all transmit powers reduces the absolute interference levels, it changes the SNRs, only making the signals more vulnerable to noise and multipath. Moreover, with less transmitting power the message requires more hops. Second, as the absolute carrier sensing threshold remains unchanged, lower transmit powers is equivalent to less sensitive carrier sensing and allowing multiple simultaneous transmission in the network.

CSMA outperforms TSCH in terms of latency, as shown in Fig. 5c. This is because the different time slots are always allocated in the same manner. For example, node 2 always transmits in the same time slot for the different transmitting powers. However, the average latency for TSCH is many times higher than the latency of CSMA/CA, and with TSCH the latency increases with an increasing density. In the simulated TSCH a slot frame contains n time slots; for a 400-node network, a slot frame has 77 time slots to cover the whole network. One time slot has duration of 10 ms; therefore, the slot frame duration is 770 ms. During this time the dimming application sends seven messages from the application layer down to the MAC layer, but the node can only send one message. The rest get stored in a queue and have to wait at least 770 ms.

Figure 5b shows that the packet loss rate (PLR), even with the smallest number of nodes in the network, is around 3 percent, but this number increases aggressively with an increasing number of nodes. With 400 nodes the PLR is between 30 and 40 percent. In fact, 30 to 40 percent fail to get the message because of excessive transmissions/retransmissions and collisions. At –25 dBm, nodes are barely or not in transmitting range of each other. If the density for –25 dBm increases, the PLR also decreases, because the nodes are getting closer to each other and the probability of being in transmission range increases. However, the phenomenon of a decreasing PLR for lower transmit powers is because of a change in the relationship between the power of the interference and the clear channel assessment (CCA) threshold. If CSMA performs CCA in a setup with lower transmit power, the power on the channel will be below the CCA threshold more often, which means CSMA will attempt to transmit instead of resuming the backoff. This creates more transmission and leads to a lower number of messages being lost by having reached the backoff retry limit. Figure 5d shows the PLR for TSCH. As expected for a TDMA-based MAC-Protocol, the PLR for 0 dBm to -10 dBm is zero. For –25 dBm the PLR decreases with an increasing density,

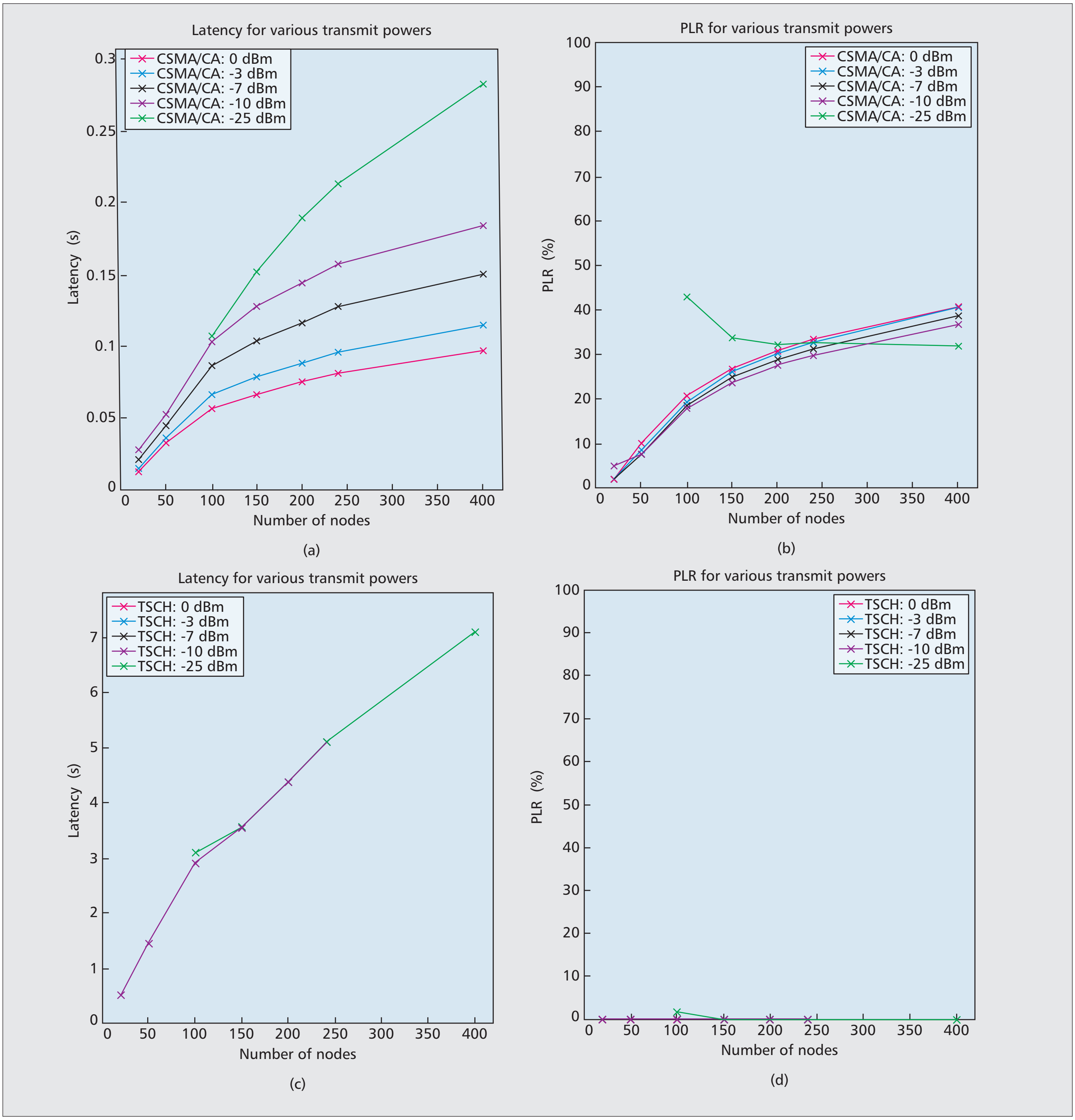


Figure 5. a) Latency with CSMA/CA for various transmit powers; b) packet loss rate with CSMA/CA; c) latency with time slotted channel hopping (TSCH); d) PLR with TSCH.

because the same problem occurs then with CSMA. The transmission range is too small, so some packets get lost in noise, and if the network density is at 200 nodes, the PLR is back at 0 percent. However, in any case when a packet gets lost in a collision or by CS, the lamp does not change the light level or turn on. At the end of turning a dimming knob, the lighting network will have many different light levels or even lamps that have not turned on. If some lamps are turning on and others are not, the user will interpret this as “a few lamps are broken.”

The simulation results show clearly that both MAC protocols combined with simple flooding are not suitable for a lighting use case. CSMA/CA might fulfil the latency requirement, but the PLR is high. On the other hand, TSCH has a 100 percent delivery rate but at the cost of latency, which is 35 times higher than the threshold of 200 ms in the worst case.

Conclusions and Outlook

Lighting control poses challenges to in-building networks that are not well met with current wireless standards. Although the exchange of room occupancies or a sequence of user commands does not create high bandwidth traffic, the latencies and outage probabilities seen in current networks would affect the user satisfaction. In particular, the avoidance of unnecessary broadcast storms is an area where further research can be impactful. Possibly improved application protocols for data sharing (e.g., Trickle) are promising. The combination of low-latency broadcast traffic and regular flow of data gathering is a further relevant area.

While LEDs promise a major reduction of energy usage in buildings, the standby power consumption of a wireless network may become significant. It would be attractive to have better schemes to allow low receiver duty cycles while maintaining latency and reachability.

The extreme density of lighting nodes leads to propagation conditions that demand more accurate modeling of the interference than only considering the nearest neighbors. This is an area in which theoretical models and simulations may require improvement to realistically predict under which conditions the network can be scaled.

The extreme density of lighting nodes leads to propagation conditions that demand more accurate modelling of the interference than only considering the nearest neighbours. This is an area in which theoretical models and simulations may require improvement to realistically predict under which conditions the network can be scaled.

Acknowledgment

The research leading to these results has received funding from the European Union's Seventh Framework Program (FP7/2007-2013) under the Artemis Joint Undertaking "DEWI–Dependable Embedded Wireless Infrastructure" project under grant agreement No 621353 and from specific national programs and/or funding authorities (e.g., Enterprise Ireland).

The authors would like to thank Dee Denteneer and Michael Kuhn for providing valuable feedback during the preparation of this article.

References

[1] D. De Guglielmo, G. Anastasi, and A. Seghetti, "From IEEE 802.15. 4 to IEEE 802.15. 4e: A Step Towards the Internet of Things," *Advances onto the Internet of Things*, vol. 260, S. Gaglio and G. Lo Re, Eds. Springer, 2014, pp. 135–52.

[2] F. Chen *et al.*, "Performance Evaluation of IEEE 802.15.4 LR-WPAN for Industrial Applications," *2008 5th Annual Conf. Wireless On Demand Network Systems and Services*, 2008, pp. 89–96.

[3] J. Wang, "Zigbee Light Link and Its Applicationss," *IEEE Wireless Commun.*, vol. 20, no. 4, Aug. 2013, pp. 6–7.

[4] F. Chen, R. German, and F. Dressler, "Towards IEEE 802.15.4e: A Study of Performance Aspects," *2010 8th IEEE Int'l. Conf. Pervasive Computing and Commun. Wksps.)*, 2010, pp. 68–73.

[5] R. Sombrutzki and A. Zubow, "A Practical Approach to Reliable Flooding in Mobile Ad hoc Networks," 2014, https://sar.informatik.hu-berlin.de/research/publications/SAR-TR-2013-06/SAR-TR-2013-06_.pdf; accessed 17 Apr., 2014].

[6] S.-Y. Ni *et al.*, "The Broadcast Storm Problem in A Mobile Ad Hoc Network," *Proc. 5th Annu. ACM/IEEE Int'l. Conf. Mob. Comput. Net. — MobiCom '99*, 1999, pp. 151–62.

[7] P. Levis *et al.*, "Trickle: A Self-Regulating Algorithm for Code Propagation and Maintenance in Wireless Sensor Networks," *Proc. 1st Symp. Networked Systems Design and Implementation*, 2004, pp. 15–28.

[8] C. Cano *et al.*, "A Low Power Listening MAC with Scheduled Wake Up After Transmissions for WSNs," *IEEE Commun. Letters*, vol. 13, no. 4, Apr. 2009, pp. 221–23.

[9] W. Ye, F. Silva, and J. Heidemann, "Ultra-Low Duty Cycle MAC with Scheduled Channel Polling," *Proc. 4th Int. Conf. Embedded Networked Sensor Systems '06*, 2006, p. 321.

[10] P. Baronti *et al.*, "Wireless Sensor Networks: A Survey on the State of the Art and the 802.15.4 and ZigBee Standards," *Comp. Commun.*, vol. 30, no. 7, May 2007, pp. 1655–95.

[11] P. Elejoste *et al.*, "An Easy to Deploy Street Light Control System Based on Wireless Communication and LED Technology," *Sensors (Basel).*, vol. 13, no. 5, Jan. 2013, pp. 6492–6523.

[12] J. M. G. Linnartz, "On the Performance of Packet-Switched Cellular Networks for Wireless Data Communications," *Wireless Networks*, vol. 1, no. 2, June 1995, pp. 129–38.

[13] N. Abramson, "The Throughput of Packet Broadcasting Channels," *IEEE Trans. Commun.*, vol. 25, no. 1, Jan. 1977, pp. 117–28.

[14] J.-P. Linnartz and D. V Perez, "Effect of Path Loss on Outage Probability in Multi-Hop Broadcast Networks," *2013 IEEE 20th Symp. Commun. and Vehic. Tech. in the Benelux*, 2013, pp. 1–6.

[15] P. Dong, "Assessment of Link Quality in 2.45GHz High-Density IEEE 802.15.4 Wireless Networks," *2013 Int'l. Symp. Intelligent Signal Proc. and Commun. Sys.*, 2013, pp. 481–86.

Biographies

Conrad Dandelski received his B.Eng. in electrical engineering from Hochschule Darmstadt — University of Applied Sciences, Germany in 2013. He is currently working on dense wireless and actuator networks for building lighting management at the Nimbus Centre for Embedded Systems Research, where he is pursuing an M.Eng. by Research degree.

Bernd-Ludwig Wenning received Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering and information technology from the University of Bremen, Germany, in 2002 and 2009. From 2002 to 2012, he was a full-time researcher at the University of Bremen, participating in several projects in the area of communication networks. In 2012, he joined the Nimbus Centre at Cork Institute of Technology, Ireland. Since then, he has been involved in a number of projects around wireless sensor networks. Throughout his research career, he has published more than 30 papers in conferences and journals.

Daniel Viramontes Perez received his M.Sc. in electrical engineering (Broadband Telecommunication Technology track) at Eindhoven University of Technology (TU/e), Netherlands. During his graduation project he worked with Philips Research. He is currently working with Vodafone Netherlands in the Network and Service Operations Department.

Dirk Pesch is head of the Nimbus Centre for Embedded Systems Research at Cork Institute of Technology, Cork, Ireland, where he has been since 1999. He received a Dipl.Ing. degree from RWTH Aachen University, Germany, and a Ph.D. from the University of Strathclyde, Glasgow, Scotland, both in electrical and electronic engineering. His research interests focus on design and performance characterization, protocols, and services for heterogeneous wireless networks, wireless sensor networks, machine-to-machine communication, and vehicular networks with applications for energy management, intelligent transportation, and smart cities. He has over 20 years of research and development experience in national and EU funded projects in both academia and industry. He has made 200 journal, book chapter, and conference publications in his area of expertise. He is also heavily involved in international conference organization in the area of wireless and mobile systems and applications.

Jean-Paul Linnartz [F '11] is a part-time professor at Eindhoven University of Technology in the Signal Processing Systems (SPS) group and a research fellow at Philips Research in Eindhoven. His current research interests focus on intelligent lighting systems. As a senior director at Philips Research, he has led research groups in the area of information security (2003–2006), wireless connectivity (2006–2008), and IC design (2009–2011). In 1992–1993, he was an assistant professor at the University of California at Berkeley. From 1988 to 1991 and in 1994, he was with Delft University of Technology, Netherlands, respectively. During 1987–1988, he worked with the the Netherlands Organization for Applied Scientific Research (TNO) on frequency planning and UHF propagation. It was for leadership in security with noisy data that he was made a Fellow of the IEEE. He holds more than 40 granted U.S. patents.

MILLIMETER-WAVE COMMUNICATIONS FOR 5G – PART 2: APPLICATIONS

Maged Elkashlan

Trung Q. Duong

Hsiao-Hwa Chen

In the September 2014 issue of *IEEE Communications Magazine*, the first part of this Feature Topic included five articles that covered the fundamentals of mmWave communications with topics ranging from propagation to coverage, presenting a holistic view of research challenges and opportunities in the emerging area of mmWave radio systems and 5G mobile broadband. The use of this technology is expected to surge in the next few years and to transform the Internet industry in the next 10 years. This part of the Feature Topic will address in more detail many technical and application issues related to beamforming, device-to-device communications, heterogeneous networks, and multimedia transmission.

We start with three articles addressing the challenging problems of coverage, initial access, handover, and hybrid beamforming. The first article, "Multi-Gigabit Millimeter Wave Wireless Communications for 5G: From Fixed Access to Cellular Networks" by Peng Wang *et al.*, explores the potential of the E-band spectrum for future mobile communications by introducing the E-band spectrum and its propagation characteristics. The authors discuss the possible applications over E-band frequencies with emphasis on the network architecture and the air interface design of E-band mobile broadband. Several key techniques are discussed that can potentially solve the coverage problem and provide good link qualities regardless of the locations of the mobile users in the network area. The second article, "Random Access in Millimeter-Wave Beamforming Cellular Networks: Issues and Approaches" by Cheol Jeong *et al.*, explores the random access channel (RACH) in mmWave communications as an important procedure for initial access and handover. Since random access cannot fully benefit from beamforming due to lack of information about the best transmit-receive beam pair, the design of RACH becomes a particularly challenging problem, especially in non-line-of-sight (NLOS) channels. The main problem of random access is that the total duration of RACH should be long to accommodate the multiple preambles transmitted for all transmit and receive beam pairs. To overcome this challenge, the authors propose several potential solutions such as enhanced preamble detection performance, multiple digital chains at the base station (BS), beam reciprocity, and cell planning. The third article, "Large-Scale Antenna System with Hybrid Analog and Digital Beamforming for Millimeter Wave 5G" by ShuangFeng Han *et al.*, explores the optimal design of hybrid analog and digital beamforming in a multi-user scenario. The authors examine the general scenario of N transceivers and M antennas per transceiver, where the energy efficiency (EE) and spectrum efficiency (SE) of the $N \times M$ beamforming structure is analyzed. Several key insights are drawn from the EE-SE relationship at the green point, the point with the highest EE on the EE-SE curve.

Cost-effective and scalable wireless backhaul solutions are therefore essential for realizing the 5G vision of anywhere anytime multi-gigabit-per-second data rates. The backhaul is clearly a fundamental component for supporting network densification in small cell deployments with very low latency inter-BS communication to combat intercell interference. In the fourth article, "Point to Multipoint In-Band mm-Wave Backhaul for 5G Networks" by Rakesh Taori *et al.*, presents an in-band solution to meet the backhaul and inter-BS coordination challenges that come with network densification. The authors argue that in-band wireless backhaul for data backhauling and inter BS-coordination is feasible without significantly hurting the cell access capacities.

Mobile broadband will continue to drive the demands for higher traffic and higher end-user data rates. These demands for very high system capacity and very high end-user data rates can be met using technologies such as ultra-dense networks, device-to-device communications, and heterogeneous networks. The next four articles discuss these important technologies by considering practical issues related to mobility, resource sharing, deployment, and routing. The fifth article, "Ultra-Dense Networks in Millimeter-Wave Frequencies" by Robert Baldemair *et al.*, explores ultra-dense networks as a solution to fulfill the extremely high traffic demands on system capacity and achievable end-user date rate. Ultra-dense networks are networks with distances between access nodes ranging from a few meters in indoor deployments up to roughly 50 m in outdoor deployments. The authors present an extensive survey detailing key requirements and characteristics of ultra-dense networks, taking into account mobility, self-backhauling, and spectrum sharing. The sixth article, "Enabling Device-to-Device Communication in Millimeter Wave 5G Cellular Networks" by Jian Qiao *et al.*, introduces a hybrid system architecture of D2D communications over mmWave 5G cellular networks. The most relevant aspects regarding the propagation characteristics to achieve reliable high-rate communications are discussed, including propagation loss, diffraction and penetration ability, and directional antennas. The authors propose an effective resource sharing scheme that allows non-interfering D2D links to operate concurrently, taking into account neighbor discovery for handoff and network integration for high-rate D2D communications with mobility. The seventh article, "Hybrid Millimeter-Wave Systems: A Novel Paradigm for HetNets" by Hani Mehrpouyan *et al.*, proposes an mmWave heterogeneous architecture, hybrid HetNet, which exploits the bandwidth and propagation characteristics of the V- and E-bands to reduce the impact of interference in HetNets. Two transceiver structures that enable handoffs from the V-band to the E-band, and vice versa are proposed. The eighth article, "10 Gb/s HetSNets with Millimeter-Wave Communications: Access and Networking Challenges and Protocols" by Kan Zheng *et al.*, introduces heterogeneous and small cell networks (HetSNets) in mmWave communications to increase spatial spectrum reusability and efficiency. The authors focus on the network and medium access control layers, taking into account several deployment scenarios for backhaul and user access. The authors discuss issues related to routing, access control, and interference coordination.

Multimedia applications such as on-demand HD video, IPTV, and ultra HDTV demand ultra-high throughput and ultra-low latency. MmWave technologies such as IEEE 802.15.3c and IEEE 802.11ad are ideal candidates to handle the data volumes in wireless multimedia data-intensive applications, in particular to multiple users demanding differentiated quality of service (QoS). MmWave multimedia communications is therefore an important topic, and is addressed in the next two articles. The ninth article, "Millimeter Wave Multimedia Communications: Challenges, Methodology, and Applications" by Dan Wu *et al.*, defines and evaluates several important metrics to characterize the multimedia QoS in mmWave communications and jointly takes into account several technical challenges, including large-scale attenuation, atmospheric absorption, phase noise, and limited gain amplifiers. The authors design a QoS-aware multimedia scheduling scheme to achieve the trade-off between performance and complexity, in which accurate propagation analysis is carried out and suitable countermeasure techniques are pointed out to satisfy the QoS requirements. The final article, "Multimedia Resource Allocation in MmWave 5G Networks" by Sandra Scott-Hayward *et al.*, presents an optimization solution, known as particle swarm optimization (PSO), as an ideal candidate to handle a mixed set of multimedia applications. The authors propose channel time allocation PSO as a reduced execution time solution to successfully optimize the channel time allocation in a mixed multimedia wireless environment, even in scenarios where blockage exists in the mmWave network.

BIOGRAPHIES

MAGED ELKASHLAN (maged.elkashlan@qmul.ac.uk) received his Ph.D. degree in electrical engineering from the University of British Columbia, Canada, in 2006. From 2006 to 2007, he was with the Laboratory for Advanced Networking at the University of British Columbia. From 2007 to 2011, he was with the Wireless and Networking Technologies Laboratory at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. He also held an adjunct appointment at the University of Technology Sydney, Australia, between 2008 and 2011. In 2011, he joined the School of Electronic Engineering and Computer Science at Queen Mary, University of London, United Kingdom, as an assistant professor. His research interests include millimeter wave communications, energy harvesting, cognitive radio, and wireless security. He currently serves as an Editor for *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, and *IEEE Communications Letters*. He received the best paper award at IEEE ICC '2014, the International Conference on Communications and Networking in China (CHINACOM) in 2014, and IEEE Conference VTC-Spring '13. He received an Exemplary Reviewer Certificate for *IEEE Communications Letters* in 2012.

TRUNG Q. DUONG (trung.q.duong@qub.ac.uk) received his Ph.D. degree in telecommunications systems at Blekinge Institute of Technology, Sweden, in 2012. He was a visiting scholar at Polytechnic Institute of New York University from December 2009 to January 2010, and at Singapore University of Technology and Design from July 2012 to August 2012. In 2013, he joined Queen's University Belfast as an assistant professor. He has served as a TPC member for major IEEE conferences, including ICC, GLOBECOM, WCNC, VTC, and PIMRC. He currently serves as an Editor for *IEEE Communications Letters* and *Wiley Transactions on Emerging Telecommunications Technologies*. He also served as Lead Guest Editor of the Special Issue on Secure Physical Layer Communications of *IET Communications*, Guest Editor of the Special Issue on Cooperative Cognitive Networks of the EURASIP *Journal on Wireless Communications and Networking*, and Guest Editor of the Special Issue on Security Challenges and Issues in Cognitive Radio Networks of the EURASIP *Journal on Advances in Signal Processing*. He received the best paper award at IEEE ICC '14, CHINACOM '14, and IEEE VTC-Spring '13. He received an Exemplary Reviewer Certificate for *IEEE Communications Letters* in 2012. His current research interests include merging green communications and networking technologies such as cross-layer design, cooperative communications, cognitive radio networks, and physical layer security.

HSIAO-HWA CHEN [F] (hshwchen@gmail.com) is currently a Distinguished Professor in the Department of Engineering Science, National Cheng Kung University, Taiwan. He obtained his B.Sc. and M.Sc. degrees from Zhejiang University, China, and a Ph.D. degree from the University of Oulu, Finland, in 1982, 1985, and 1991, respectively. He has authored or co-authored over 400 technical papers in major international journals and conferences, six books, and more than 10 book chapters in the communications area. He is Editor-in-Chief of *IEEE Wireless Communications*. He has served as General Chair, TPC Chair, and Symposium Chair for many international conferences. He is the founding Editor-in-Chief of Wiley's *Security and Communications Networks Journal*. He was the recipient of the best paper award at IEEE WCNC 2008 and the IEEE Radio Communications Committee Outstanding Service Award in 2008. He is a Fellow of IET and BCS.

Multi-Gigabit Millimeter Wave Wireless Communications for 5G: From Fixed Access to Cellular Networks

Peng Wang, Yonghui Li, Lingyang Song, and Branka Vucetic

Peng Wang, Yonghui Li, and Branka Vucetic are with the University of Sydney.

Lingyang Song is with Peking University.

ABSTRACT

With the formidable growth of various booming wireless communication services that require ever increasing data throughputs, the conventional microwave band below 10 GHz, which is currently used by almost all mobile communication systems, is going to reach its saturation point within just a few years. Therefore, the attention of radio system designers has been pushed toward ever higher segments of the frequency spectrum in a quest for increased capacity. In this article we investigate the feasibility, advantages, and challenges of future wireless communications over the E-band frequencies. We start with a brief review of the history of the E-band spectrum and its light licensing policy as well as benefits/challenges. Then we introduce the propagation characteristics of E-band signals, based on which some potential fixed and mobile applications at the E-band are investigated. In particular, we analyze the achievability of a nontrivial multiplexing gain in fixed point-to-point E-band links, and propose an E-band mobile broadband (EMB) system as a candidate for the next generation mobile communication networks. The channelization and frame structure of the EMB system are discussed in detail.

INTRODUCTION

In recent years video on demand, videoconferencing, online gaming, e-education, and e-health have been introduced to a rapidly growing population of global subscribers using devices such as laptops, tablets, and smartphones. The formidable growth in demand for these communication services requires ever increasing data throughputs. To cater to this growing demand, many advanced technologies have been adopted in the current fourth-generation (4G) systems, such as Long Term Evolution (LTE) and Mobile WiMAX, to substantially increase the transmission rate. These technologies, including orthogonal frequency-division multiplexing (OFDM), multiple-input multiple-output (MIMO), multi-user detection, advanced channel coding (e.g. turbo and low-density parity-check, LDPC, coding), adaptive coding and modulation, hybrid automatic repeat request (HARQ), cell splitting, and heterogeneous networking have made the achievable spectrum efficiency very close to the theoretical limits. Existing cellular systems all operate below 10 GHz frequency bands that are already heavily utilized. Therefore, there is little space to further increase the transmission rate in these frequency bands. The attention of radio system designers has been pushed toward ever higher segments of the frequency spectrum in a quest for capacity increase.

The millimeter wave (mmWave) band from 30 to 300 GHz offers large swathes of spectrum [1], potentially forming the basis for the next revolution in wireless communications. As predicted in [2], after excluding some subbands with severe atmospheric absorption and assuming 40 percent of the remaining spectrum potentially becoming available over time, a possible 100 GHz new spectrum among the mmWave band could be opened up for future mobile communication use. However, this is an optimistic forecast as this possible 100 GHz spectrum is discretely distributed in the overall mmWave band with distinct channel characteristics and various service restrictions imposed by regulators in different countries. Uniting these discrete segments of bandwidths collectively for mobile broadband communication use will remain a great challenge. Comparatively, the frequency bands 71–76 GHz and 81–86 GHz, collectively called the E-band, have been released by the International Tele- communication Union (ITU) to provide broadband wireless services [3]. Different from the severe oxygen absorption in the 60 GHz band, which contributes about 15 dB/km of attenuation in addition to free space losses, atmospheric absorption above 70 GHz drops significantly to less than 1 dB/km and rises again after 100 GHz due to molecular effects [4]. Therefore, the E-band opens a large frequency window with low atmospheric attenuation, making it very suitable for long-distance wireless transmissions. This 10 GHz spectrum in the E-band, which is about 50 times the bandwidth of the

0163-6804/15/$25.00 © 2015 IEEE

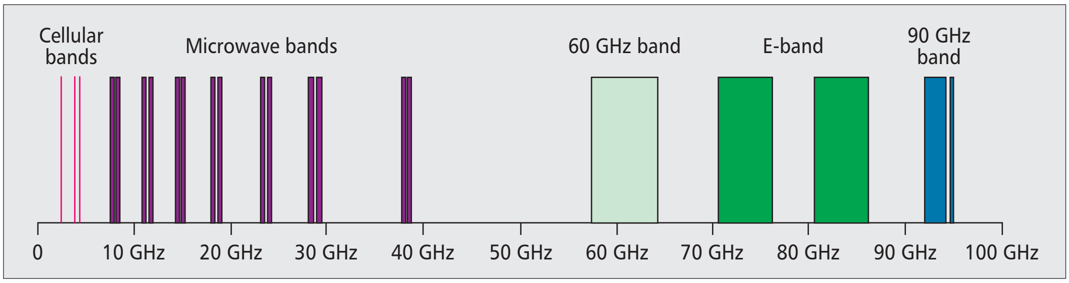


Figure 1. E-band frequency allocation.

entire current cellular spectrum, is by far the widest ever allocated by the Federal Communications Commission (FCC) at any one time, and can provide 5 GHz bandwidth per channel for accommodating multi-gigabits per second and even higher data rates with greatly reduced latency over large distances.

There have already been some commercial E-band wireless systems for fixed point-to-point communications. For example, by utilizing the leading-edge radio frequency (RF) monolithic microwave integrated circuit (MMIC) technology [5], the E-link 1000 G1 radio from the E-band Communications Corporation can provide best-in-class E-band link performance for gigabits-per-second data rates over a distance of up to a few kilometers. It has been forecast that in the near future the fifth generation (5G) of cellular communication systems will be developed over untapped mmWave bands. The superior propagation characteristics of E-band frequencies make this band preferable over the other segments of mmWave bands. Although E-band transceivers are presented with new design challenges such as increased phase noise, limited amplifier gain, and the need for transmission line modeling of circuit components, the electronics industry is rapidly developing, producing component electronics with ever reducing physical sizes and power consumption. This means the hardware preparation for a mobile communication system over E-band will be ready. The combination of cost-effective complementary metal oxide semiconductor (CMOS) technology and high-gain steerable antennas at the devices and base stations (BSs) will strengthen the viability of E-band communications.

In this article we discuss the potential of exploring the E-band spectrum for future mobile communications. We first present a brief review of the history of E-band spectrum and its light licensing policy as well as benefits/challenges. Then we introduce the propagation characteristics of E-band signals, based on which some potential fixed and mobile applications at the E-band are investigated. In particular, we analyze the achievability of nontrivial multiplexing gain in fixed point-to-point E-band links and propose an E-band mobile broadband (EMB) system as a candidate for the next generation mobile communications. The channelization and frame architecture of the EMB system are discussed in detail. Finally, we conclude the article with a brief summary.

E-band Spectrum

A Brief History of E-band

The E-band allocations for fixed services were first established by the ITU at the 1979 WARC-79 World Radio Communication Conference. However, not much commercial interest was shown in this band until the late 1990s, when the FCC published a study on the use of themillimeter-wave bands [4]. Afterward, the FCC made a historic ruling in 2002 to open up the E-band for exclusive federal governmental use in the United States. A novel “light licensing” scheme was introduced in 2005 [6] and the first commercial E-band radios were installed soon after. Canada adopted the same bands with the same technical specifications and licensing regimens as the United States in 2005. Also, in 2005 the European Conference for Postal and Telecommunications Administrations (CEPT) released a Europe-wide band plan for fixed services in the E-band, which was modified lat in 2009. In 2006 the European Telecommunications Standards Institute (ETSI) released technical rules for equipment operating in the E-band. Similar specifications are also effective or proposed for the United Kingdom and Australia. Nowadays many parts of the world have followed the United States and European leads, and have opened up the E-band frequencies for enabling gigabit-per-second speed point-to-point wireless transmissions.

E-band Frequency Allocation

The E-band frequency allocation consists of the two unchannelized bands of 71–76 GHz and 81–86 GHz,[1] as shown in Fig. 1. This combined 10 GHz of spectrum is significantly larger than any other frequency allocation, enabling a whole new generation of wireless transmission to be realized. In addition, different from the lower microwave frequency bands, which are sliced into subchannels of no more than 50 MHz, which in turn limits the data rate transmitted over them, the E-band spectrum is only divided into a pair of 5 GHz channels and not further partitioned. These two 5 GHz channels at E-band are 100 times the size of even the largest microwave channel. Such an unpartitioned spectrum allocation allows us to support gigabit-per-second data rates for each signal using relatively simple system architectures and modulation schemes. Radio equipment can take advantage of low-order modulation modems, nonlinear power amplifiers, low-cost diplexers, direct conversion receivers, and many more relatively non-complex wireless building blocks, leading to reduced sys-

Radio equipment can take advantage of low order modulation modems, nonlinear power amplifiers, low-cost diplexers, direct conversion receivers, and many more relatively non-complex wireless building blocks, leading to reduced system cost and complexity while increasing reliability and overall radio performance.

[1] *In the United States and Canada, the E-band spectrum also includes 92–95 GHz except 94–94.1 GHz.*

Country	License structure	License fee
United States	Online light license	$75 for 10-year license
United Kingdom	Light license	£50 per year
Russia	Light license	Minimal registration fee
Australia	Light license	AU$187 per year

Table 1. Typical E-band license structures and license fees in some countries.

tem cost and complexity with increased reliability and overall radio performance.

Light Licensing for E-Band

In many countries the conventional license application of microwave bands from regulators requires a long period measured in months or even years. The corresponding licensing fee is determined by formulas depending on either the transmission data rate, the required bandwidth, or both. Such formulas can result in prohibitively high license fees for high-capacity systems. To promote E-band commercialization, the national wireless link regulators and administrators in many countries have introduced innovative and streamlined "light licensing" [7] schemes for managing this band, providing the E-band with an attractive alternative to existing licensed frequency bands. The "light licensing" policy allows the E-band licenses to be applied for in minutes and at a cost of a few tens of dollars per year, significantly faster and cheaper than traditional licensing. Thousands of fixed E-band radios have been registered and installed in these countries. The typical E-band license fees in several countries are listed in Table 1.

The "light licensing" policy comes from three unique characteristics of the E-band. First, since there are very few E-band services currently, it is argued that the spectrum at E-band frequencies is no longer scarce. Second, the high frequencies at E-band allow the systems to adopt highly directional antennas and communicate via highly focused "pencil beam" transmissions, leading to dense configuration of communication links without interference concern and thus a high degree of frequency reuse. Third, the E-band frequencies are configured as a single pair of 5 GHz channels, which makes the traditional frequency planning/coordination unnecessary and the related interference analysis significantly simplified. Thus, the E-band administration and cost of license are dramatically reduced. The "light licensing" at the E-band reflects the ease of coordinating, registering, and licensing, and sets license fees that cover administrative costs, but does not penalize the high data rates and bandwidths that are required for ultra-broadband services. It is worth noting that despite the name "light licensing," the possession of such a license still gives the link operator the same full benefits of a traditional link license, including link registration, "first come first served" rights, and full interference protection.

Benefits of E-Band over Other Wireless Technologies

There are many technologies available to provide wireless broadband connectivity and fiber-like services. These technologies include WiFi, 60 GHz wireless, free space optics (FSO), and so on. E-band wireless systems offer significant benefits over them with the following advantages [8].

High antenna gains and allowable output power: Thanks to the small wavelength of E-band signals, it is possible to realize large gains from relatively small antennas at E-band frequencies. In addition, the FCC permits E-band radios to operate with up to 3 W of output power, significantly higher than that available at other mmWave bands (e.g., 25 dB higher than the 10 mW limit at 60 GHz). The high antenna gain and high output power allow E-band radios to overcome the higher rain fading and foliage losses experienced at E-band frequencies.

Guaranteed high data rates: E-band offers much higher data rates, e.g. Gb/s and above, than any other wireless technology. Such high rates are guaranteed even under deteriorated transmission conditions such as rain, which beats WiFi, WiMAX, and other broad-coverage technologies whose system performance depends heavily on the radio and user environments.

Long distance transmissions with robust weather resilience: E-band wireless allows Gb/s-level transmission over a very long distance up to 12 miles, much longer than those supporting similar data rate such as 60 GHz and FSO systems. This long-distance transmission is robust to almost all environmental conditions such as fog, dust, air turbulence, and other atmospheric impairment that can disable optical links for hours.

Low-cost and rapid licensing policy providing guaranteed interference protection: Under the "light licensing" policy, licenses for E-band links can be obtained much faster and cheaper than those for traditional microwave bands, and in the meanwhile provide the full benefits of traditional link licenses that grant full interference protection from other nearby wireless sources. Even in the unlikely event of interference, the full weight of the wireless regulator is available to identify and remove the interference source.

Cost-effective fiber-like wireless solution: The cost of high-capacity E-band wireless systems is only a fraction of that of buried fiber alternatives. Installed wireless systems have payback periods of months compared to the costs of trenching new fiber. Installing dedicated wireless technology can often be more economical than leasing fiber-provided high-capacity services.

A summary of the most important system parameters and network characteristics of various broadband techniques are detailed in Table 2.

Technical Research Challenges of E-Band Communications

Although numerous benefits are presented above, there are still some challenging technical issues that must be addressed before commercialization of the E-band frequencies. They include:

	WiFi	3/4G	60 GHz	FSO	Fiber	E-band
Data rate	About 1 Mb/s, unstable	About 10 Mb/s, unstable	100∼1000 Mb/s	100∼1000 Mb/s	Up to 100s of Gb/s	Mutiple Gb/s
Transmission distance	20 yards	2 miles	500 yards	200 yards	Up to 60 miles	Up to 12 miles
Licensing	free for unlicensed use	Licensed spectrum very scarce	Free for unlicensed use	Not regulated	N/A	Light license
License cost	N/A	High	N/A	N/A	N/A	Low
License application period	N/A	Months/years	N/A	N/A	N/A	Minutes/hours
Guaranteed interference protection	No	Yes	No	No	Yes	Yes
Installation time	Hours	Months/years	Hours/days	Hours/days	Months/years	Hours/days
Installation cost	Low	High	Medium	Medium	High	Medium

Table 2. Comparison of different broadband techniques.

- Severe E-band propagation loss.
- Unclear channel modeling at such high frequencies.
- High transceiver complexity in such large MIMO systems with a massive number of antennas.
- Hardware constraint imposed to E-band transceiver design, where a large number of antennas have to be driven by a limited number of radio-frequency (RF) chains due to the high cost and power consumption of the latter.

However, as detailed in the rest of this article, all these issues can potentially be, and are already being, effectively addressed. The severe propagation loss can readily be compensated through deploying a large number of transmit/receive antennas that provide significant beamforming gains. Several research groups have already conducted E-band propagation measurements in real urban environments [9], providing some fundamental hints for the proper modeling of E-band channels. Some initial and efficient channel estimation algorithms [10] that utilize the channel sparsity, and a hybrid precoder design [11] that relieves the high hardware costs, have also been proposed. All these developments have made E-band a very promising candidate frequency segment for future 5G wireless broadband mobile communications.

E-Band Propagation

While signals at lower frequency bands propagate for several tens of miles and penetrate easily through buildings, E-band signals can travel only a few miles or less, and do not penetrate solid materials very well. However, these characteristics are not necessarily disadvantageous. In fact, the propagation loss can be exploited to reduce interference, increase frequency reuse, and prevent eavesdropping, thus providing very efficient spectrum utilization and increasing the security of communication transmissions.

Free Space Propagation

Due to the small wavelength of E-band signals, transmissions over the E-band are principally contributed by line-of-sight (LoS) components. According to the free space transmission model, the path gain of the LoS link between two omnidirectional antennas with distance D is mathematically expressed as

$$G = G_T G_R \frac{\lambda^2}{(4\pi D)^2} \qquad (1)$$

where G_T and G_R are, respectively, the gains of transmit and receive antennas, and λ is the signal wavelength. It is seen from Eq. 1 that given G_R, G_T, and D, the path gain is proportional to λ^2, indicating that the E-band transmissions suffer much more power loss than those over conventional microwave bands. For example, the propagation at 75 GHz is 30 dB worse than that at 2.4 GHz (the operating frequency for WiFi networks). Thus, to guarantee the same signal power (and in turn the same quality of service) at the receiver, the transmitted power at 75 GHz must be 30 dB higher than that at 2.4 GHz. This makes the signal transmission/reception through a single omnidirectional antenna practically infeasible in E-band systems.

One approach to compensate for the severe E-band power loss is to equip a massive number of antennas at both link ends to provide a large beamforming gain. Different from conventional microwave systems where the large-size antennas must be sufficiently spaced and may lead to extraordinarily large transmitter/receiver aperture sizes, this approach can easily be implemented in E-band systems as the antenna size and spacing scale down with the wavelength. The synthesized low-cost antenna arrays can be electronically steered to provide adaptive yet highly directional links, permitting a flexible deployment. In principle, the number of antenna elements that can be packed into a given aperture size is increased by four times for every doubling

With well understood information on rainfall characteristics in particular regions, it is easy to design E-band ratio links capable of overcoming the worst weather conditions via adaptive transmit power control, or predict the levels of weather outage of longer links.

of the operating frequency, providing about 6 dB beamforming gain at each link end if these antennas are compactly located to form an equivalent directional antenna for steering a "pencil beam." When the beamforming gains at both link ends are taken into consideration, the overall power gain of the link then scales as $1/\lambda^2$. Therefore, the propagation at 75 GHz becomes 30 dB better than that at 2.4 GHz, which implies a significant redeeming feature of multiple antenna transmissions in the E-band.

Blockage, Multipath, and Scattering

Pure free space propagation between the transmitter and receiver happens only when the LoS component is present and no building/obstacle is around. In practice, an E-band communication link is always located within a building group area, in which the buildings, cars, and even human beings may either block the LoS transmission or "bend" the signal impinging on their surfaces. The corresponding propagation characterizations of E-band signals are very different from those of traditional microwave ones. Due to the small wavelength on the order of several millimeters, transmissions over E-band are effectively blocked by obstacles such as wooden boards and brick walls. In addition, E-band signals are also not prone to diffraction when encountering an obstacle, which is similar to light waves. Reflection constitutes the most received signal power among all non-LoS (NLoS) links. Principally, the signal power received from each reflected link may be much lower than that from a LoS link. This is because, besides partial absorption by reflecting materials, E-band signals encounter greater diffusion and less specular reflection than microwave signals due to the relatively "rougher" reflecting material surface compared to their signal. Even though, it has been experimentally validated [9] that these NLoS links can still provide substantial link connection and coverage extension in mmWave cellular systems, especially when LoS transmissions are unavailable. According to the E-band propagation measurements conducted by NYU WIRELESS in the dense urban environment of New York City [9], the path loss exponent for NLoS propagation is 5.88 with a shadow factor of 14.19 dB, which is a result of several different paths of great dynamic range supported over a wide range of angles. Therefore, although the large buildings on every city block and crowded streets cause numerous blockages, they also create reflections and scatters between the transmitter and receiver with slightly more path loss and fewer multipath components than those measured at a lower frequency of 28 GHz. This indicates that E-band transmissions will be able to rely on multipath environments and directional antennas to overcome additional propagation loss at E-band. In addition, the smaller number of multipath components relative to those of the transmit/receive antennas endure the sparse nature of the E-band propagation channel, which may significantly reduce the operational complexity involved in channel estimation and transceiver design. Provided that the angle of departure (AoD) and angle of arrival (AoA) information of each path is available, we can combine multipath components with different AoDs/AoAs to significantly improve the path loss exponents and link margins through beamforming and beam combining. This will make it feasible to deploy a mobile communication network over E-band with reasonable BS coverage and acceptable outage performance.

Other Attenuation Factors at E-Band

In addition to the power loss during the free space and reflected/scattered propagations, transmission over E-band also suffers from some other attenuation factors, as detailed below.

Atmospheric attenuation: When traveling through the atmosphere, the E-band signals may be absorbed by molecules of oxygen, water vapor, and other gaseous atmospheric constituents. Fortunately, these losses are merely about 0.5 dB/km in total, much less than those at 60 and 100 GHz above, and close to that of the popular microwave frequencies. This makes the E-band frequencies very favorable for radio transmissions over many miles under clear conditions.

Fog and clouds: Since fog and cloud particles are much smaller than the E-band wavelengths, the attenuation caused by them is almost negligible, for example, only an attenuation of 0.4 dB/km is led by thick fog at density of 0.1 g/m3 (about a visibility of 50 m). Comparatively, the attenuation for an FSO optical signal caused by heavy fog could be about 200 dB/km due to the similar magnitudes of the signal wavelength and fog/cloud particles.

Dust and other small particles: Similar to fog and cloud particles, the magnitudes of these particles are much smaller than the E-band wavelengths, making them essentially invisible to E-band transmissions.

Rain: Transmissions at E-band experience significant attenuation in the presence of rain [12], which places practical limits on the link distances. For example, "heavy" rainfall at the rate of 25 mm/h can lead to over 10 dB/km attenuation at E-band frequencies. The corresponding attenuation even reaches up to 30 dB/km in the case of tropical rainfall with a rate of 100 mm/h. Fortunately, most intensive rain tends to fall in limited parts of the world, mainly the Equatorial countries. In other countries such as the United States, Canada, and Australia, such severe weather generally occurs only in very short bursts. It tends to fall in small and dense clusters within a larger and lower-intensity rain cloud, and is usually associated with a severe weather event that moves quickly across the link path. Therefore, rain outage tends to be short and is only problematic on longer-distance transmissions. With well understood information on rainfall characteristics in particular regions, it is easy to design E-band ratio links capable of overcoming the worst weather conditions via adaptive transmit power control, or predict the levels of weather outage of longer links.

Ice crystals and snow: Ice crystals and snow do not cause appreciable attenuation, even if the rate of fall exceeds 125 mm/h. This is due to the much reduced loss of ice compared to water.

Foliage: Foliage losses are significant at E-band frequencies and may be a limiting propagation impairment for E-band transmissions.

For example, the foliage loss at 75 GHz for a penetration of 8 m (roughly equal to the diameter of a large tree) is about 20 dB.

In summary, E-band propagations exhibit comparable characteristics to those at the widely used microwave bands, and with well characterized weather characteristics allowing rain fade to be understood, link distances of several miles can confidently be realized.

Fixed E-Band Applications

A wide range of fixed services are realizable over E-band frequencies. The following are some examples.

Last-mile access: In many communities, the last-mile access technique represents a major remaining challenge because the cost of providing high-speed high-bandwidth services to individual subscribers in remote areas can be higher than the service provider would like. Laying wire and fiber optic cables is an expensive undertaking that can be environmentally demanding and require high maintenance. Many experts believe that broadband wireless networks will eventually solve this difficulty and meet everyone's needs. E-band frequencies provide a promising solution in terms of flexibility, speed, and cost of construction.

Wireless backhaul: With the rapid growth of mobile data traffic, traditional backhaul that utilizes narrow bandwidth is consequently regarded as a potential bottleneck for the overall cellular system. E-band offers a cost-effective and flexible alternative to fiber for future backhaul. Access points and BSs can easily be connected via E-band links, providing gigabit-per-second backhaul capacity to solve this bottleneck problem.

Network recovery: In the case of fiber breakage, a fixed point-to-point E-band link can be used to provide temporary service restoration, due to its much shorter setup time and lower cost in comparison with those required to restore the original fiber link.

Campus LAN: Fixed E-band links can also be installed to directly build up a gigabit wireless LAN within a building group (e.g., campus) as the extension of fiber optic communication networks. High-speed gigabit access will be maintained within both the wireless and wired parts of the overall communication network, but without the problems and expenses related to fiber installation.

Storage access: Machine-to-machine connectivity for storage area networks could easily be established via fixed point-to-point E-band links with excellent data security and high availability.

In all the above applications, both terminals of a communication link are usually fixedly located (e.g., on the tops or side walls of buildings) such that LoS transmissions are guaranteed. Therefore, the corresponding channel is mainly contributed by LoS transmissions, and other effects such as multipath, foliage loss, and atmospheric stratification are not significant due to the extremely narrow beams in which the radiation propagates. The primary sources of link impairments come from adjacent link interference, rain attenuation, and antenna perturbation. The adjacent link interference occurs when the LoS power of one link is directed into the main lobe or a side-lobe of the receive antennas in adjacent links. This impairment can be avoided in advance via proper organization of link deployments. Rain attenuation is well understood, and the resultant impairment can be compensated via adaptive transmit power control. The antenna perturbation, potentially caused by wind induced pole sway or other environmental concerns, may lead to severe mismatch between the directivities of transmit/receive antennas. Some robust and computationally efficient beam alignment technique may be required to combat this problem.

The antenna perturbation, potentially incurred by wind-induced pole sway or other environmental concerns, may lead to severe mismatch between the directivities of transmit/receive antennas. Some robust and computationally efficient beam alignment technique may be required to combat this problem.

A main concern for fixed E-band systems is whether a high-capacity link can be guaranteed between the transmitter and receiver to support gigabits or even tens-of-gigabits throughputs over a given distance. As previously mentioned, multiple antenna techniques are essential in the E-band to provide beamforming gain to compensate for the severe propagation loss. To further enhance the link capacity, we need to rely on multiple antenna techniques to achieve a multiplexing gain such that transmissions of multiple spatially independent signal streams can be supported simultaneously without interfering with each other. However, this is not a trivial problem in the E-band systems. Unlike rich-scattering microwave channels where independent Rayleigh faded coefficients between different antenna links enable the achievable multiplexing gain to increase linearly with the minimum number of transmit/receive antennas, the fixed E-band channels are dominated by nearly deterministic LoS components, and the achievable multiplexing gain heavily depends on the antenna deployments at both link ends.

In spite of this, it has been shown [13, 14] that the maximum multiplexing gain is still principally achievable in fixed E-band channels, provided that the geometrical distributions of the antennas at both link ends are carefully designed. For example, in an E-band LoS MIMO channel with aligned uniform linear antenna arrays (ULAs) at both the transmitter and receiver, the maximum multiplexing gain is achieved when the following *Rayleigh distance criterion* is fulfilled [13]

$$D = D_{Ray} = \frac{\max\{N_t, N_r\} d_t d_r}{\lambda} \qquad (2)$$

where N_t (N_r) and d_t (d_r) are, respectively, the number and spacing of the transmit (receive) antennas. In this case, the channel contains $\min\{N_t, N_r\}$ eigenmodes with equal channel gains, indicating that the maximum multiplexing gain, $\min\{N_t, N_r\}$, is indeed achievable, and thus that many spatially independent signal streams can be supported simultaneously. Similar observations have also been made in the general situation when the ULAs at both ends have arbitrary orientations [14]. However, the antenna spacings, d_t and d_r, in a practical E-band LoS MIMO system may be limited by the physical sizes of the transmitter/receiver and cannot be arbitrarily large. Consequently, the communication distance that satisfies the Rayleigh distance criterion is also limited, indicating that the maximum multiplexing gain is not always achievable in practice.

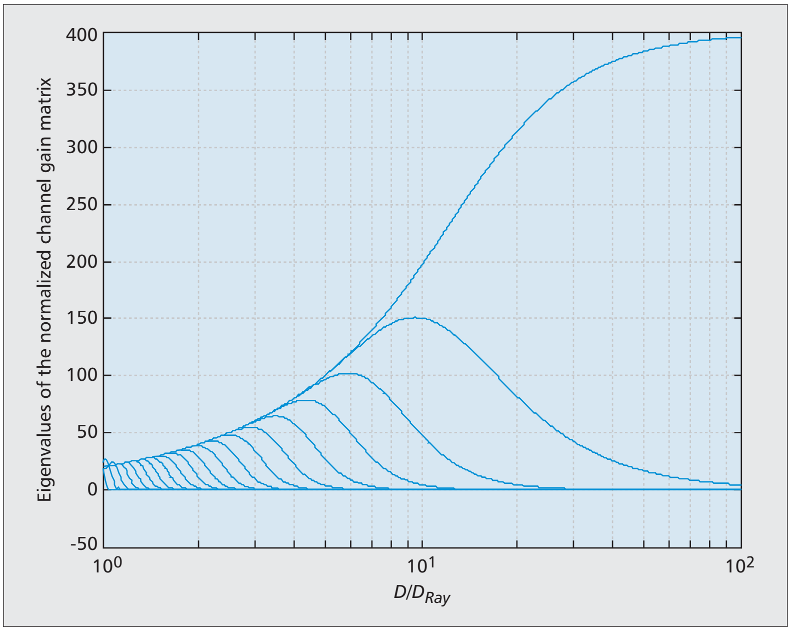


Figure 2. The eigenvalue curves for the normalized channel gain matrix of a ULA-based E-band LoS MIMO channel with 20 antennas at both the transmitter and receiver (i.e., $N_t = N_r = 20$).

Figure 2 shows all the eigenvalue curves of an aligned ULA-based E-band LoS MIMO channel with 20 antennas at both ends (i.e., $N_t = N_r = 20$). In Fig. 2 we assume a far-field distance between the two link ends such that the channel coefficients for all antenna links have approximately the same amplitude, which is normalized to 1 for convenience. It is seen that for the E-band system beyond the Rayleigh distance (i.e., $D > D_{Ray}$), although the channel may still be of full rank, some of its eigenmodes are very poor, and signal transmissions over them will be very inefficient in practice. For convenience, we denote by $\mu_m(D)$ the mth largest eigenvalue of a ULA-based E-band LoS MIMO channel gain matrix and account the m^{th} eigenmode as an effective eigenmode if $\mu_m(D)/\mu_1(D) \geq \gamma$, where γ is a threshold related to the system working signal-to-noise ratio (SNR). As a consequence, the number of effective eigenmodes can be referred to as the effective degree of freedom (EDOF) of the channel. It is shown in [15] that when N_t and N_r are sufficiently large, the farthest distance that can provide an EDOF of m (and in turn can support m spatially independent signal streams) is mathematically given by

$$D_{\max}^{(m)} = c_m(\gamma)\frac{N_t d_t N_r d_r}{\lambda}$$
$$\approx c_m(\gamma)\frac{(N_t-1)d_t(N_r-1)d_r}{\lambda} = c_m(\gamma)\frac{D_t D_r}{\lambda} \quad (3)$$

where $D_{\max}^{(m)}$ is referred to as the maximum effective multiplexing distance of EDOF-m, $D_t = (N_t - 1)d_t$ and $D_r = (N_r - 1)d_r$ are the aperture sizes of the transmit and receive ULAs, respectively, and $c_m(\gamma)$ is a constant function. Equation 3 indicates that the farthest distance that can support a given number of spatially independent signal streams at a finite SNR is mainly determined by the product of the aperture sizes of the transmit/receive ULAs, instead of the numbers of antennas at both ends. Hence, to support a higher number of spatially independent signal streams in a ULA-based E-band LoS MIMO channel, we must either increase the product of transmit/ receive aperture sizes or reduce the communication distance.

E-Band Mobile Broadband Communications

In this section we discuss the feasibility and challenge of establishing an E-band mobile broadband (EMB) network. As mentioned earlier, E-band signals do not penetrate solid materials very well. This implies that the overall EMB networks can be effectively isolated into indoor and outdoor networks by the brick walls of buildings. For indoor networks, plenty of reflecting materials are present, making NLoS transmissions (also called diffuse links) very common in such scenarios. Therefore, indoor mobile users can easily access the network via access points installed in each room without suffering from weather impairment. The Doppler effect is not a concern either as the relatively small indoor serving area restricts user mobility. Since this scenario has been extensively investigated for communication over other segments of mmWave bands such as 60 GHz, we do not discuss it here. In what follows, we assume that handoff between indoor and outdoor networks is guaranteed via the access points equipped at the entrances of the buildings and mainly focus on the outdoor networks.

A common myth in the wireless engineering community is that rain and foliage attenuation make E-band spectrum practically useless for outdoor mobile communications. However, the outdoor EMB network can overcome these issues and provide a seamless user experience after adopting the following potential techniques.

Dense EMB BS Deployment

To guarantee a reasonably high probability of successful link connection between the BSs and mobile users and provide sufficiently good coverage, it is preferable to equip BSs densely in a given EMB network area so as to combat both the severe path attenuation experienced by E-band signals and the possible block of LoS transmissions caused by surrounding buildings/obstacles. The BS antennas could be located adaptively according to the topography and architectural construction of the serving area, e.g. on the surface of buildings or the top of lampposts along the streets and at each street corner. The E-band propagation measurements conducted by NYU WIRELESS in the dense urban environment of New York City [9] have revealed that for inter-site distances up to 200 meters, atmospheric attenuation is of a negligible degree and the rain attenuation is only about 2 dB for a heavy rainfall of 25 mm/hr. Therefore, a cell size on the order of about 200 meters, similar to today's microcell sizes, is sufficient to guarantee qualified LoS links in urban environments. Thanks to the distinctive narrow beam technique adopted at E-band, the interference among adjacent EMB BSs can be significantly supressed and thus their coverage areas can be largely overlapped.

Adaptive Beamforming

Beamforming is another efficient technique to overcome path attenuation. At the transmitter, the signal is emitted from different antennas with different phases and amplitudes, creating constructive or destructive patterns at intended or undesired receivers. At the receiver, signals from different receive antennas are combined together using a set of weight coefficients such that the power or SNR of the collected signal after combination is maximized. When an LoS link is available between the mobile user and the BS, proper beamforming/combining patterns that point to each other can be generated at both link ends so as to significantly enhance the link quality. On the other hand, when an LoS link is unavailable, adaptive beamforming is still capable of enhancing the NLoS link quality by exploiting multipath in urban environments. In this case, the surrounding buildings, especially those with smooth surfaces made of glass or marble, could provide stronger reflection and less diffusion. Thus, the signal transmission and reception can be directed to such strong reflected NLoS links using adaptive beamforming. Satisfactory link quality can still be achieved together with proper adaptive transmit power control.

Sparse Channel Estimation

To explore the potential benefit of adaptive beamforming, accurate and timely channel state information (CSI) is crucial in an EMB network. Recall that a massive number of antennas are necessarily required at one or both link ends to provide sufficient power gain in compensating the severe E-band propagation loss. This indicates a significant increase of CSI overhead to be estimated at the receiver and fed back to the transmitter. Fortunately, recent research results [9] have revealed that due to the much higher E-band signal frequencies, an E-band channel generally consists of a much smaller number of paths between the transmitter and receiver than its antenna numbers equipped at both link ends, even in the dense urban environment. This indicates that an E-band channel can exhibit a sparse nature after being converted into the beam-space domain, and by utilizing this sparse property, the CSI overhead in an EMB network can be significantly reduced. A channel estimation algorithm that explores this sparsity in EMB networks has already been proposed [10], which directly works on the sparse version of the channel matrix after it has been converted into the beam-space domain and can quickly estimate the AoDs/AoA and fading coefficient of each path in a bi-section searching manner. More efficient and advanced channel estimation approaches are also under investigation.

Hybrid Tranceiver Design

Due to user mobility, the beamforming/combining vectors need to be adaptively adjusted so that the beams are always pointing to each other as the mobile user moves. However, different from the adaptive beamforming technique for traditional microwave that can be implemented digitally at baseband, the adaptive beamforming design in E-band is restricted by the hardware constraint, where a large number of antennas are driven by a limited length of RF chain due to the high cost and power consumption of the latter. Hybrid digital-and-analog precoder/combiner design is a practical solution to this difficulty [11]. With hybrid precoding/combining, the transmitters/receivers are able to apply a high-dimensional RF precoder, implemented via analog phase shifters, followed by a low-dimensional digital precoder that can be implemented at baseband. Near optimal unconstrained performance can be achieved at practically low cost.

With hybrid precoding/combining, the transmitters/receivers are able to apply a high-dimensional RF precoder, implemented via analog phase shifters, followed by a low-dimensional digital precoder that can be implemented at baseband. Near optimal unconstrained performance can be achieved at practically low cost.

User Cooperation

When the surrounding buildings of an E-band network have relatively rough surfaces, the reflected signal power may be much reduced, and the link quality cannot be guaranteed if the LoS link between the mobile user and BS is blocked. User cooperation may provide a solution to this situation. Specifically, we can build up certain reward mechanisms to encourage vacant users with good-quality links to BSs to serve as relays and help forward data for other users with bad link qualities. The overall network may work as follows. First, all the EMB BSs continuously broadcast their pilot signals selected from a pilot set $\mathcal{P}_1$ through a signaling channel. The pilot signals used by different BSs are referred to as level-1 pilots and assumed to be mutually orthogonal. Each mobile user in the serving area, whether it has data to transmit/receive or not, estimates the qualities of the links to different BSs based on the received level-1 pilot signals. These mobile users are then classified into directly served (DS) users and indirectly served (IS) users, according to the link qualities to the surrounding BSs. A DS user refers to a mobile user that has at least one BS to which the link quality is better than a certain threshold. Contrarily, an IS user refers to the user whose link qualities to all BSs are below the threshold. The operations for DS and IS users are different and introduced separately below.

Operation for DS users: Each DS user chooses the BS with the best link quality as its serving BS and registers to its serving BS for future data transmission, reception, or forwarding. If a DS user has data to transmit/receive, a traffic channel is assigned by the serving BS for performing data transmission/reception in a similar way to the operation in traditional cellular networks. Otherwise, if this DS user is vacant, it will keep listening to the channel and, in the meanwhile, broadcast the link quality information between it and the serving BS together with a pilot signal (referred to as level-2 pilot) selected from another pilot set $\mathcal{P}_2$. By this means and assuming that all the pilots in $\mathcal{P}_1$ and $\mathcal{P}_2$ are mutually orthogonal, these vacant DS users will serve as potential relays to help the data forwarding of IS users.

Operation for IS users: After failing to connect to any BS due to lack of both LoS and strong NLoS reflected links, an IS user will measure the qualities of the links to all the surrounding vacant DS users who are broadcasting

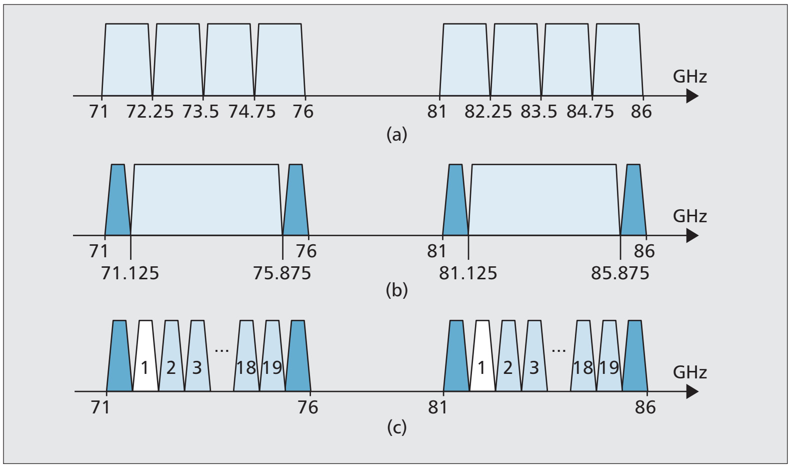


Figure 3. E-band channelization in: a) the United States and Canada; b) the United Kingdom and Australia; c) Europe.

level-2 pilots, based on which the overall qualities of all the possible IS user-vacant DS user-BS links are calculated. Afterward, this IS user sends a request to the vacant DS user through which the overall two-hop link quality is the best. Upon receiving and accepting this request, the selected vacant DS user then builds up an indirect link between its serving BS and the served IS user, helping the latter to register to the former indirectly. When data transmission/reception is required by this IS user, the selected DS user applies for a traffic channel fofrom its serving BS, enabling the data of this IS user to be forwarded to/from the core network.

Hybrid EMB and 4G Systems

Due to the nature of E-band signal propagation, it is possible that even with the abovementioned techniques, some dead spots still exist in which the mobile users cannot be served by the EMB network. Therefore, a mechanism that supports emergency communications when the communications over the E-band are not successful should be considered as part of the EMB system design. A hybrid EMB and the current 4G cellular network infrastructure may be adopted to provide better coverage and seamless user experience, as well as preserve the benefit of gigabit transmissions at E-band. In such a hybrid network, the overall E-band frequencies are mostly utilized for data transmission. The system information, control channel, and feedback channel are implemented over the current 4G frequency band. In addition, some 4G cellular frequencies should be preserved for data transmission of the mobile users at dead spots.

Channelization and Frame Structure

Although ITU has released the E-band frequencies for fixed and mobile services, there has been no specific recommendation regarding the use of the E-band or sharing arrangements with other services. Nevertheless, many countries have issued their E-band channelization plans to promote its commercialization. Figure 3 summarizes the E-band channelization plans in several representative areas. In the United States and Canada, both the 71–76 and 81–86 GHz bands are divided into four unpaired 1.25 GHz segments (eight in total) without mandating specific channels within them, and these segments may be aggregated without limit. In Europe, the United Kingdom, and Australia, a 125 MHz guard band is set at the top and bottom of each 5 GHz sub-band of the E-band spectrum to prevent potential interference to and from adjacent bands. In particular, the United Kingdom and Australia have no explicit channel plan for the remaining segments of the E-band, while Europe further divides each of the two 4.75 GHz bands into nineteen 250 MHz channels and allows aggregation of any number of channels from 1 to 19. Furthermore, the specified channels may be used for either time-division duplex (TDD) or frequency-division duplex (FDD) systems either within the single band or in combination with other bands.

Here we propose a possible frame structure for EMB systems based on the European channelization plan. Note that since the European channelization plan is compatible with that of the United Kingdom and Australia, our proposed frame structure is also applicable to the latter two countries. Following the current 4G systems, we choose OFDM as the multiplexing scheme for EMB due to its superiority in efficient multiple access and simpler equalization at the receiver. As shown in Fig. 4, the durations of one frame and subframe are chosen to be 10 ms and 1 ms, respectively, which are the same as those of LTE systems in order to facilitate hybrid EMB and 4G operation.

The other parameters in the OFDM numerology are designed as follows. Since the bandwidth of each channel in the European channelization plan is 250 MHz, we choose the sampling rate as 30.72 MHz $\times$ 8 = 245.76 MHz, where 30.72 MHz is a popular frequency at which a good trade-off can be achieved between clock accuracy and cost. In addition, we select the subcarrier spacing to be 480 kHz based on the following reasons:

- First, from the implementation viewpoint, the fast/inverse fast Fourier transform (FFT/IFFT) size, denoted by K, is typically a power of 2, meaning that the subcarrier spacing should have a form of 30.72×2^{-k} MHz for some integer k. The value of 480 KHz satisfies this form when $k = 8$.
- Second, due to the high directional transmission characteristic of EMB, the corresponding maximum delay spread may be limited to a few nanoseconds, which in turn leads to a much wider coherent bandwidth than that in LTE. Accordingly, the subcarrier spacing of 480 kHz is small enough to stay within the coherent bandwidth of most situations in EMB.
- Third, by assuming that the moving speed of mobile users is no more than 120 km/h, the resultant Doppler shift, f_d, is at most 120 km/h $\times$ 86 GHz/(3 $\times$ 10^8m/s) $\approx$ 10 kHz. This value is much less than 480 kHz and thus can keep intercarrier interference due to Doppler sufficiently low.

- Fourth, with a reasonable clock accuracy of 10 ppm, the corresponding clock drift at E-band is at most 10 ppm × 86 GHz = 860 kHz, which should be less than two times the subcarrier spacing to enable simple system synchronization and acquisition.
- Finally, the 480 kHz subcarrier bandwidth indicates an FFT/IFFT size of 512 points for the overall 250 MHz bandwidth of each channel, which is small enough in complexity because this size takes about 20 percent of the RX digital baseband complexity.

Furthermore, since the channel coherent time is $T_c = 1/f_d \approx 0.1$ ms determined by the above calculated Doppler shift, we divide each subframe into 32 slots such that each slot has a duration of 31.25 µs, which is less than the channel coherent time. The number of OFDM symbols in each slot is set to 14 with the corresponding cyclic prefix (CP) lengths being about 0.179 µs (44 samples) for the first OFDM symbol and 0.146 µs (36 samples) for the remaining 13 OFDM symbols. Such a design leads to a CP overhead of about 6.7 percent and provides sufficient margin to cope with the maximum delay spread and synchronization error.

Conclusions

In this article we have introduced the background and propagation characteristics of E-band transmissions. In particular, the potential of exploring the E-band spectrum for mobile broadband communications in the coming few decades is discussed. E-band transmissions rely heavily on directional beamforming with very narrow beam widths, allowing effective suppression of interference among adjacent E-band mobile broadband BSs and significant overlap of their coverage areas. Also, because of directional beamforming, a key challenge in the E-band mobile broadband network is to guarantee good coverage of the overall network, especially when some mobile users do not have LoS links to the surrounding BSs. Several techniques have been discussed that can potentially solve the coverage problem and provide good link qualities regardless of the locations of the mobile users in the network area. A hybrid EMB and 4G system may provide a good trade-off between the coverage and data rate.

Acknowledgment

The authors would like to thank Mr. Yongping Zhang, Ms. Jiahui Qiu, and Ms. Miao Wang for enlightening discussions on the design of EMB frame structure.

References

[1] D. Lockie and D. Peck, "High-Data-Rate Millimeter-Wave Radios," *IEEE Microwave Mag.*, vol. 10, no. 5, Aug. 2009, pp. 75–83.

[2] Z. Pi and F. Khan, "An Introduction to Millimeter-Wave Mobile Broadband Systems," *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 101–07.

[3] FCC 03-248, "Allocation and Service Rules for the 71–76 GHz, 81–86 GHz and 92–95 GHz Bands," Nov. 2003.

[4] FCC Bulletin 70, "Millimeter Wave Propagation: Spectrum Management Implications," July 1997.

[5] R. A. Pucel, "Looking Back at Monolithic Microwave Integrated Circuits," *IEEE Microwave Mag.*, vol. 13, no. 4, May 2012, pp. 62–76.

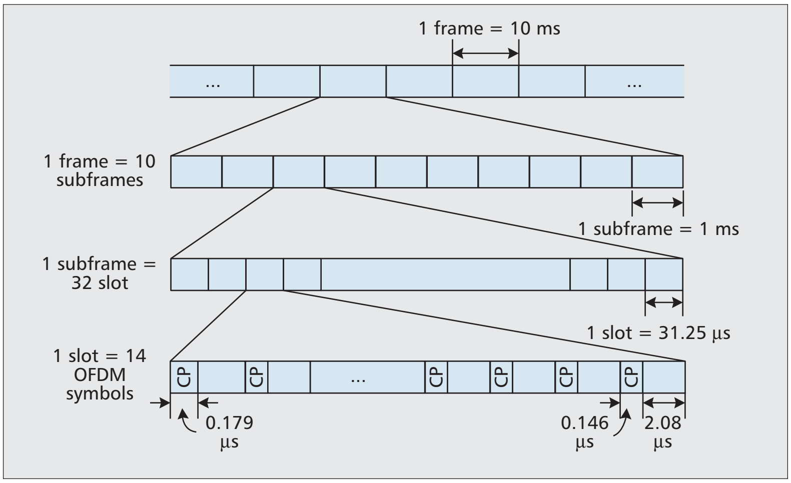


Figure 4. Frame structure of the EMB system.

[6] FCC Report and Order, "Allocations and Service Rules for the 71–76 GHz, 81–86 GHz, and 92–95 GHz Bands," 05-45, 2005.

[7] E-Band Communications Corp., "Light Licensing Benefits of the 71–76 & 81–86 GHz Frequency Bands," http://www.e-band.com.

[8] E-Band Communications Corp., "Benefits of E-Band Systems over Other Solutions," at http://www.e-band.com.

[9] G. R. MacCartney Jr. and T. S. Rappaport, "73 GHz Millimeter Wave Propagation Measurements for Outdoor Urban Mobile and Backhaul Communications in New York City," *Proc. IEEE ICC '14*, Sydney, Australia, 10–14 June 2014.

[10] A. Alkhateeb *et al.*, "Channel Estimation and Hybrid Precoding for Millimetre Wave Cellular Systems," *IEEE J. Sel. Topics Signal Processing*, vol. 8, no. 5, May 2014, pp. 831–46.

[11] O. El Ayach *et al.*, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, May 2013, pp. 1499–1513.

[12] ITU-R P.838-3, "Specific Attenuation Model for Rain for Use in Prediction Methods," 2005.

[13] D. Gesbert *et al.*, "Outdoor MIMO Wireless Channels: Models and performance Predication," *IEEE Trans. Commun.*, vol. 50, no. 12, Dec. 2002, pp. 1926–34.

[14] F. Bohagen, P. Orten, and G. E. Oien, "Design of Optimal High-Rank Line-of-Sight MIMO Channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, Apr. 2007, pp. 1420–24.

[15] P. Wang *et al.*, "Tens of Gigabits Wireless Communications over E-band LoS MIMO Channels with Uniform Linear Arrays," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, July 2014, pp. 3791–3805.

Biographies

Peng Wang [S'05, M'10] received his B.Eng. degree in telecommunication engineering and M.Eng. degree in information engineering from Xidian University, Xi'an, China, in 2001 and 2004, respectively, and his Ph.D. in electronic engineering from the City University of Hong Kong in 2010. He was a research fellow with the City University of Hong Kong and a visiting postdoctoral research fellow with the Chinese University of Hong Kong from 2010 to 2012. Since 2012 he has been with the Centre of Excellence in Telecommunications, School of Electrical and Information Engineering, University of Sydney, Australia, where he is currently a research fellow. His research interests include channel and network coding, information theory, iterative multi-user detection, MIMO techniques, and millimeter-wave communications. He has published more than 40 peer-reviewed research papers in leading international journals and conferences, and has served on a number of technical programs for international conferences such as ICC and WCNC.

Yonghui Li (M'04, SM'09) received his Ph.D. in November 2002 from Beijing University of Aeronautics and Astronautics. From 1999 to 2003 he was affiliated with Linkair Communication Inc., where he held the position of project manager with responsibility for the design of physical layer solutions for the LAS-CDMA system. Since 2003 he has been with the Centre of Excellence in Telecommunications at the University of Sydney. He is now an associate professor in the School of Electrical and Information Engineering, University of Sydney. He was the Australian Queen Elizabeth II Fellow and is currently the Australian Future Fellow. His current research interests are in the area of wireless communications, with a particular focus on MIMO, cooperative communications, coding techniques, and wireless sensor networks. He holds a number of patents granted and pending in these fields. He is an Executive Editor for *European Transactions on Telecommunications* (ETT). He has also been involved in the technical committees of several international conferences, such as ICC and GLOBECOM.

Lingyang Song [S’03, M’06, SM’12[received his Ph.D. from the University of York, United Kingdom, in 2007, where he received the K. M. Stott Prize for excellent research. He worked as a postdoctoral research fellow at the University of Oslo, Norway, and Harvard University, until rejoining Philips Research UK in March 2008. In May 2009 he joined the School of Electronics Engineering and Computer Science, Peking University, China, as a full professor. His main research interests include MIMO, OFDM, cooperative communications, cognitive radio, physical layer security, game theory, and wireless ad hoc/sensor networks. He has received best paper awards at many conferences, including the IEEE International Conference on Wireless Communications, Networking and Mobile Computing, the First IEEE International Conference on Communications in China, the 7th International Conference on Communications and Networking in China, the IEEE WCNC ’12, the International Conference on Wireless Communications and Signal Processing, and IEEE ICC ’14. He is currently on the Editorial Boards of *IEEE Transactions on Wireless Communications*, *IET Communications*, and the *Journal of Network and Computer Applications*. He is the recipient of the 2012 IEEE Asia Pacific (AP) Young Researcher Award.

Branka Vucetic [F’03] received her B.S.E.E., M.S.E.E., and Ph.D. degrees in electrical engineering from the University of Belgrade, Yugoslavia, in 1972, 1978, and 1982, respectively. She currently holds the Peter Nicol Russel Chair of Telecommunications Engineering at the University of Sydney. During her career she has held various research and academic positions in Yugoslavia, Australia, and the United Kingdom. Her research interests include wireless communications, coding, digital communication theory, and MIMO systems. She has co-authored four books and more than 300 papers in telecommunications journals and conference proceedings. She was elected to the grade of IEEE Fellow for contributions to the theory and applications of channel coding.

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 topic areas:

Analysis of new topic areas for standardization, either enhancements to existing standards, or of a new area. The standards activity may be just starting or nearing completion. For example, current topics of interest include:

- 5G radio access
- Wireless LAN
- SDN
- Ethernet
- Media codecs
- Cloud computing

Tutorials on, analysis of, and comparisons of IEEE and non-IEEE standards. For example, possible topics of interest include:

- Optical transport
- Radio access
- Power line carrier

The relationship between innovation and standardization, including, but not limited to:

- Patent policies, intellectual property rights, and antitrust law
- Examples and case studies of different kinds of innovation processes, analytical models of innovation, and new innovation methods

Technology governance aspects of standards focusing on both the socio-economic impact as well as the policies that guide it. This would include, but are not limited to:

- The national, regional, and global impacts of standards on industry, society, and economies
- The processes and organizations for creation and diffusion of standards, including the roles of organizations such as IEEE and IEEE-SA
- National and international policies and regulation for standards
- Standards and developing countries

The history of standardization, including, but not limited to:

- The cultures of different SDOs
- Standards education and its impact
- Corporate standards strategies
- The impact of open source on standards
- The impact of technology development and convergence on standards

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

Compatibility and interoperability, including testing methodologies and certification to standards

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

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

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

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

Select "Standards Supplement" from the dropdown menu of submission options.

Random Access in Millimeter-Wave Beamforming Cellular Networks: Issues and Approaches

Cheol Jeong, Jeongho Park, and Hyunkyu Yu

The authors are with Samsung Electronics Co., Ltd.

ABSTRACT

The mmWave band has been utilized for indoor applications in IEEE 802.11ad and for point-to-point wireless backhaul solutions. Very recently, mmWave communications has come into the spotlight as an enabling technology for 5G *cellular networks* by virtue of development of mmWave beamforming technology and channel measurement campaigns in outdoor environments driven by academia and industry. A high path loss in the mmWave band can be alleviated by adaptive beamforming using antenna arrays; aligning transmit and receive beams in direction can result in high beamforming gains. When we consider the cellular network in the mmWave band, random access, which is primarily used for initial access and handover, is the very first issue in system design. Since random access cannot fully benefit from beamforming due to the lack of information on the best transmit-receive beam pair, the design of the random access channel, RACH, becomes more challenging, especially in non-line-of-sight channels. In this article, we analyze fundamental issues of RACH in mmWave cellular communications and present possible approaches to address these issues. Furthermore, research challenges and future directions are discussed.

INTRODUCTION

In recent years, mobile data traffic has been dramatically increased and is expected to continue its growth. To cope with the growing demand on mobile traffic, a radio access technology needs to be enhanced by improving spectral efficiency, increasing frequency bandwidth, or increasing cell density. However, the spectral efficiency of point-to-point communication is close to the theoretical limit. Cell densification has difficulty handling a large amount of intercell interference since cells may be deployed in an unplanned manner, and the signaling for interference management is limited by non-ideal backhaul links of networks. Taking the aforementioned challenges into account, a straightforward way to deal with the traffic demand is to increase bandwidth for communications. In this sense, the millimeter-wave (mmWave) band is recently being considered as one of promising bands for cellular networks [1] since abundant contiguous frequency resources are available, while the frequency bands under 5 GHz are very fragmented and crowded.

Among the differences of propagation characteristics between legacy cellular bands under 5 GHz and mmWave bands (e.g., 28, 38, 60, and 70 GHz), the most notable one is the path loss difference. When isotropic antennas are considered at the transmitter and receiver sides, the path loss in free space increases by the frequency squared. Moreover, the penetration loss at mmWave bands is larger than that at legacy cellular bands. On the other hand, a large number of antenna elements can be packed into a small form factor in mmWave bands due to the much smaller wavelength than legacy cellular bands. Hence, the severe path loss of the mmWave bands can effectively be alleviated because a high beamforming gain is obtained using a large number of antenna elements. Moreover, a massive multiple-input multiple-output (MIMO) system [2], where hundreds of antennas are implemented at a base station (BS), becomes more feasible and can be useful for transmitting signals over a long distance in the outdoor environment. In [3], a beamforming algorithm for mmWave cellular communications is presented. The results of extensive propagation measurement campaigns in indoor and outdoor environments are shown in [4, 5]. By virtue of the beamforming technique and the propagation measurement results, the mmWave cellular network is being considered as a candidate technology for fifth generation (5G) mobile broadband.

To fully exploit the beamforming gain, the beam direction should be well aligned with the direction of the propagation path. The conventional way to find the best beam direction is to utilize the reference signal for beam quality estimation and exchange the information about the best beam direction between a BS and a mobile station (MS), as in IEEE 802.11ad protocol. This method will work well when the MS has a connection with its serving BS; that is, control signals conveying the beam index can be transmitted through an established radio link.

0163-6804/15/$25.00 © 2015 IEEE

However, there are circumstances when communication using the best beam pair may not be feasible. More specifically, when the MS accesses the network initially using a random access channel (RACH), the best beam pair cannot be known a priori. Hence, it is difficult to set the beam directions at both the MS and BS for RACH preamble transmission and reception, respectively. Since the RACH is also used for important procedures such as handover in cellular networks, the RACH design is a critical issue that needs to be resolved first.

In this article, the random access procedure in traditional cellular networks is reviewed, and the important issues in designing RACH for mmWave cellular networks are analyzed. To address those issues, possible approaches are presented, and research challenges are discussed. The random access issue in mmWave beamforming cellular systems, to the best of our knowledge, has not been thoroughly investigated in the literature yet.

Random Access in Traditional Cellular Networks

In this section, we review random access in traditional cellular networks to better understand the related issues of random access in mmWave cellular networks. In a cellular network, an MS needs to establish a radio link with a BS for data transmission and reception. To establish the radio link, the MS first acquires synchronization in downlink and then accesses the network using the RACH, as in Long Term Evolution (LTE) standard [6]. The RACH can be used for various purposes: *initial access*, *handover*, *maintaining uplink synchronization*, and *scheduling request*. Among these various purposes, we focus on initial access and handover. In general, the random access procedure consists of four steps, as illustrated in Fig. 1. In the first step, the MS randomly selects one among a set of preamble signatures. The selected preamble is transmitted from the MS to the BS using the time-frequency resources indicated in the system information broadcast by the BS. In the second step, if the BS successfully detects the random access preamble, the BS transmits the random access response (RAR), which includes the index of the detected preamble sequence, the uplink timing information, and the indication of the resource allocation for the next step. In the third step, the random access message, including the identity of the MS, is transmitted from the MS to the BS. Finally, the identity of the MS is transmitted from the BS to the MS, confirming that the random access procedure is successfully completed for the MS. In the remainder of this article, we focus on the first step in which the preamble is transmitted from the MS and detected at the BS.

MmWave Beamforming Cellular Networks

In an mmWave cellular network, highly directional beamforming is used at both the BS and MS, unlike the traditional cellular network. By virtue of very small wavelengths ranging from 1 to 10 mm in the mmWave bands, a large number of antenna elements can be packed into a small device. However, due to the high cost and complexity of hardware implementation, it is feasible to have only one or a few digital chains, each connected to a set of antenna elements that forms an analog beam. The transmitter selects a transmit beam pattern, which determines the phase shifter weights to steer the beam in a certain direction. Similarly, the receiver selects a receive beam pattern to receive the signals in a certain direction. To obtain a high beamforming gain, transmit and receive beam directions should be well aligned with each other. When the MS is in the connected state, the index of the best transmit beam of the BS is fed back from the MS to the BS periodically using the uplink control channels, and the best transmit beam of the MS can be reported through the downlink control channels so that the data can be transmitted using the best beam pair in downlink and uplink.

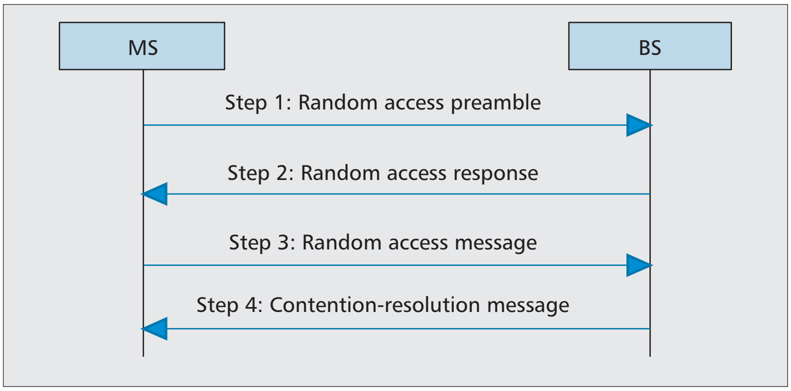

MS
BS
Step 1: Random access preamble
Step 2: Random access response
Step 3: Random access message
Step 4: Contention-resolution message

Figure 1. The random access procedure in legacy cellular networks.

The candidate frequency bands appropriate for 5G mmWave communications would be around 30 and 60 GHz. Since the free-space path loss and oxygen absorption at the 60 GHz band is larger than at the 30 GHz band, the signal attenuation at 60 GHz is larger than at 30 GHz. Hence, 60 GHz may be more appropriate for mmWave indoor communications in which line of sight (LOS) or short distance transmission are dominant, while 30 GHz may be more suitable for mmWave outdoor communications in which non-LOS (NLOS) or relatively long distance transmission should be supported.

Random Access in MmWave Cellular Networks: Omnidirectional Antenna vs. Directional Antenna

There are some cases in which the best direction cannot be previously known at either the BS or the MS. Those cases are as follows:

- When the MS tries to access the network initially
- When the MS recovers from a radio link failure (RLF)
- When the MS performs a handover procedure

There is a single MS that transmits the preambles during one RACH frame. The power delay profiles (PDPs) for multiple received signals are summed. The signature is then detected by searching the peak in the summed PDP and comparing with the predetermined threshold.

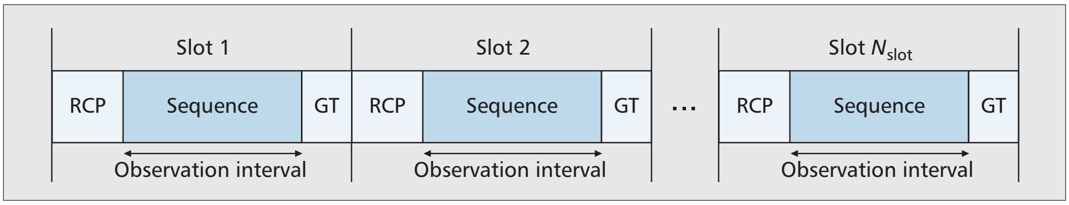


Figure 2. The frame structure for random access in mmWave beamforming cellular networks.

In these cases, the random access procedure is commonly performed between the MS and the BS. Since the best beam pair is not known, the MS has no choice but to transmit the RACH preambles in multiple directions; thus, only a few of those transmissions can achieve a high beamforming gain when the transmit and receive beams are almost aligned. With this in mind, one might ask whether beamforming is still useful although transmit and receive beams are not well aligned yet. In this section, we try to answer that question in part by comparing RACH performance between two cases with and without beamforming, and analyzing the reasons for performance difference.

An exemplary frame structure for the random access preamble transmission (i.e., the first step) is shown in Fig. 2. We simply extend the legacy RACH frame for a single preamble to the case of multiple preambles. The RACH frame consists of consecutive N_{slot} RACH slots. Each RACH slot consists of the RACH cyclic prefix (RCP), the preamble sequence, and the guard time (GT) [6]. For example, if there are $N_{\text{slot}} = M_{\text{BS}} \times M_{\text{MS}}$ RACH slots in a RACH frame where M_{BS} is the number of receive beams at the BS and M_{MS} is the number of transmit beams at the MS, the following procedure can be performed. When the first transmit beam of the MS is used for M_{BS} preamble transmissions, the BS can receive the signals using M_{BS} different receive beams. This is repeated M_{MS} times by changing the transmit beam of the MS. Using all the received signals, preamble sequence detection is performed at the BS.

Evaluation Methodology

There are two important performance metrics that are commonly used to evaluate the performance of RACH: the false alarm probability and the miss detection probability. The false alarm probability is the probability that a preamble sequence is detected when the preamble signatures are not transmitted from the MS during the corresponding RACH frame in Fig. 2. The miss detection probability is the probability that a preamble sequence is not detected when the preamble signature is transmitted in each RACH slot in the corresponding RACH frame. We evaluate the performance of RACH in terms of the miss detection probability when the false alarm probability is set to 0.1 percent. To analyze the effects of beam steering at the transmitter and the receiver, the IMT-Advanced channel model, which is a geometry-based stochastic model, is adopted [7]. The NLOS channel condition is assumed in the urban microcellular (UMi) environment. The mobile velocity is set to 3 km/h. The cell is divided into three sectors, and a BS covers one sector. Uniform linear arrays are used at both the transmitter and the receiver. There is a single MS that transmits the preambles during one RACH frame. The power delay profiles (PDPs) for multiple received signals are summed. The signature is then detected by searching the peak in the summed PDP and comparing with the predetermined threshold.

Comparison of RACH Performance

A single digital chain is used at the BS and MS, respectively. The miss detection probability is plotted vs. signal-to-noise ratio (SNR) with different numbers of antenna elements and different numbers of beams. The number of RACH slots in a RACH frame is $N_{\text{slot}} = 16$. Within the RACH frame, the whole transmit-receive beam sweeping is repeated $N_{\text{slot}}/(M_{\text{MS}}M_{\text{BS}})$ times for fair comparison; that is, the total random access duration is the same for each case. In Fig. 3, it is seen that increasing the number of antenna elements at either the transmitter or receiver provides a large performance gain compared to the case of an omnidirectional antenna (i.e., $N_{\text{MS}} = N_{BS} = 1$). We can conclude that a performance gain of the directional antenna over the omnidirectional antenna can still be achieved even if a preamble is not always transmitted and received with the best beam pair.

Array and Diversity Gains with Beamforming in Multiple Directions

The performance gains of beamforming in multiple directions compared to the omnidirectional antenna are twofold: the *array gain* and the *diversity gain*. In this subsection, it is assumed that the number of antenna elements at the BS is equal to one (i.e., $N_{\text{BS}} = 1$) without loss of generality. In the case of independent fading channels of antenna elements, there is no array gain of multiple transmit antenna elements over a single transmit antenna element (i.e., $N_{\text{MS}} = 1$). On the other hand, if the fading is fully correlated (i.e., the correlation matrix of the channel is the rank-one matrix), the average received SNR will ideally be N_{MS} times larger than the independent fading case. Since the antenna elements will be closely spaced in mmWave communications, there will be some array gain for partially correlated channels. In addition, the multiple preambles are transmitted in different directions during a RACH frame, thereby providing multiple signal branches at the receiver. Thus, the diversity gain can be achieved. Therefore, one can still obtain both the array gain and diversity gain from beamforming even if the preamble transmissions are performed in multiple directions. Similarly, in [8], it was observed that the array and diversity gains can be obtained when the MS receives the signals from multiple BSs using beamforming in a cellular network, although the article is not specifically related to random access.

Critical Issues in Random Access in Mmwave Cellular Networks

As shown in the previous section, it is better to use the beamforming technique for random access preamble transmissions using directional antennas. Based on this observation, a random access procedure in mmWave cellular networks is illustrated in Fig. 4. In the first step, preambles are transmitted repeatedly in multiple directions at the MS and received in multiple directions at the BS. The information about the best transmit beam index at the BS should also be conveyed in this step so that the best transmit beam can be used at the BS in the next step. The RAR is transmitted from the BS using the best transmit beam and received at the MS using the best receive beam in the second step. Similarly, the third and fourth steps are performed using the best transmit-receive beam pair.

This naïve approach for preamble transmission in the first step may cause a problem. A high beamforming gain can be achieved only for a few of all transmit and receive beam pairs. Most preamble transmissions cannot obtain a high beamforming gain due to the misalignment of transmit and receive beam directions. Therefore, the preamble duration for RACH should be much longer than that of other uplink control and data channels that use the best beam pair in order to achieve target coverage (e.g., a cell radius of a few hundred meters). Considering that the path loss difference between legacy bands under 5 GHz and mmWave bands around 30 GHz may be more than 20 dB in a UMi environment assuming path loss models in [7], the total duration of a RACH frame is expected to be a few tens of milliseconds, which is much longer than 1 ms in LTE. Hence, it will have a great impact on the initial access, RLF recovery, handover, uplink-downlink configuration, and beam scheduling. In the following, we discuss each issue in more detail. It is worth noting that the access method may not be a problem in IEEE 802.11ad because the service range of the protocol is very short, a few tens of meters at most.

Impacts of Long Duration of the Random Access Preamble

We identify the important issues in the mmWave cellular network by scrutinizing the impacts of four different aspects.

Initial Access and RLF Recovery — For initial access (i.e., moving from idle mode to connected mode), the random access procedure is performed. In order to reduce power consumption, an MS should spend as much time as possible in the idle state. This means that the transition time from idle state to connected state should be short. Due to the long duration of random access, however, it is not easy to design an efficient procedure related to state transition. In addition to the initial access, when the RLF occurs, the MS re-establishes the connection using random access. As the random access duration is long, the service interruption time may also be long, and thus the user experience may be heavily affected.

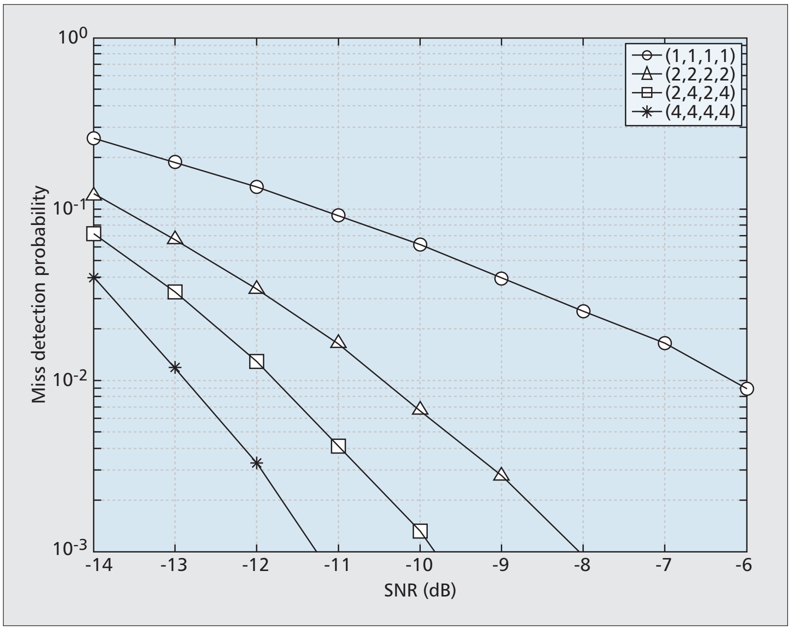


Figure 3. The miss detection probability vs. SNR where $N_{\text{slot}} = 16$. The symbol (N_{MS},N_{BS},M_{MS},M_{BS}) denotes an antenna and beam configuration where N_{MS} and N_{BS} are the numbers of antenna elements at the MS and BS, respectively; and M_{MS} and M_{BS} are the numbers of beams at the MS and BS, respectively.

Handover — There will be a large number of small cells in mmWave cellular networks. As an MS moves across cells, the handover for the MS may occur frequently. Thus, handover time should be very short for guaranteeing the quality of experience (QoE) of MSs in the network. However, because the time required for handover completion may be long due to the random access procedure, the QoE of users may be severely degraded, especially for real-time services such as voice over IP (VoIP). Although a non-contention-based random access procedure can be performed in handover (i.e., a random access preamble signature is assigned by the BS), the fact that the MS should transmit the multiple preambles in multiple directions is the same as the case of the contention-based random access procedure in mmWave cellular networks.

Uplink-Downlink Configuration — In mmWave systems, time-division duplex (TDD) may be more appropriate than frequency-division duplex (FDD) since radio resource efficiency and spectrum flexibility are important in such a wideband system. Due to the long RACH frame, however, many uplink time slots are needed. Hence, it is not flexible to configure the uplink-downlink ratio in the TDD system according to the uplink-downlink traffic ratio.

Beam Scheduling — The number of receive beams that can be simultaneously used in a time slot is limited by the number of receive digital chains at the BS. If the number of receive digital chains is small, the beam scheduling flexibility for data will be low. For example, when there is only one receive digital chain at the BS, it has no choice but to schedule an MS in a time slot where the best receive beam at the BS for the MS is the same as that used for receiving the RACH preamble in the time slot. Moreover, the problem becomes more serious when the uplink data signals for different MSs are multiplexed in the frequency domain rather in the time domain.

POSSIBLE APPROACHES TO ADDRESS THE ISSUES

As mentioned earlier, the main problem of random access is that the total duration of the RACH should be very long since multiple preambles should be transmitted for all transmit and receive beam pairs. To overcome this challenge, we present candidate solutions in the following subsections. One might think that RACH performance would be improved if the preamble bandwidth gets wider. However, we would like to note that RACH performance does not depend much on the preamble bandwidth basically given a preamble duration [6].

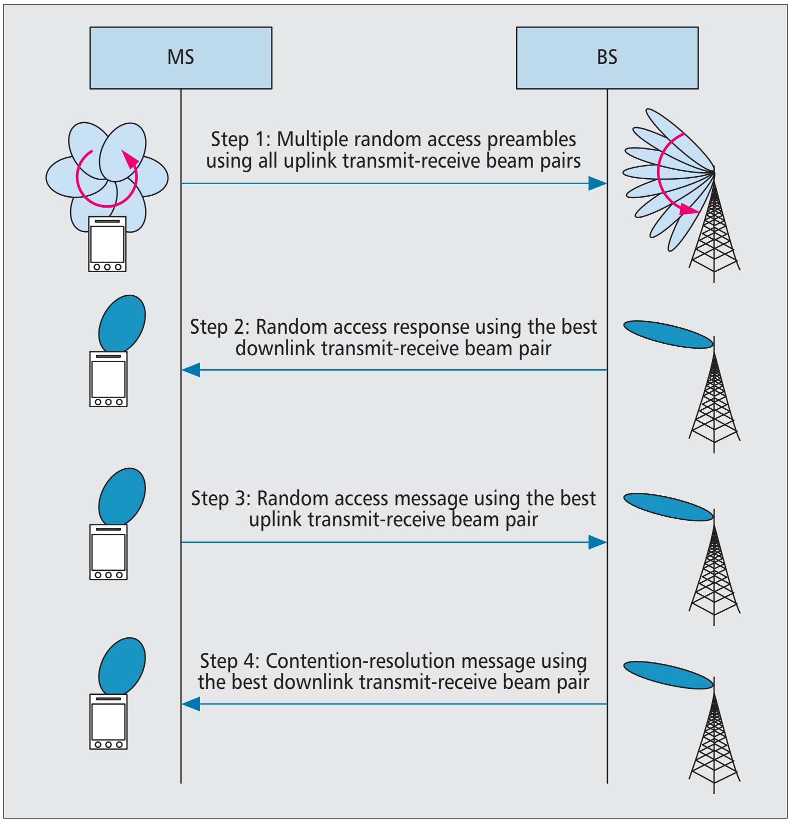

MS
BS
Step 1: Multiple random access preambles using all uplink transmit-receive beam pairs
Step 2: Random access response using the best downlink transmit-receive beam pair
Step 3: Random access message using the best uplink transmit-receive beam pair
Step 4: Contention-resolution message using the best downlink transmit-receive beam pair

Figure 4. The random access procedure in mmWave beamforming cellular networks.

ENHANCED PREAMBLE DETECTION PERFORMANCE

If we improve the performance of preamble detection, the RACH duration can be reduced for a given cell coverage. The performance depends on the preamble sequence design and the preamble detection algorithm at the receiver. It may not be easy to improve the performance by designing a new preamble sequence since the Zadoff-Chu (ZC) sequence [9] used in LTE already has desired properties such as ideal cyclic autocorrelation and minimum cross-correlation. However, one can improve the performance by enhancing the detection algorithm. For example, the conventional detection algorithm is based on the assumption that the channel response is nearly flat over RACH subcarriers. In practice, however, it is more likely that the MS experiences frequency-selective fading channels. From this point of view, an enhanced detection algorithm was proposed in [10] by taking into account the frequency selectivity of the channel. In mmWave systems, multiple signals are received at the BS; thus, the performance will depend on how those signals are combined. Therefore, a novel combining algorithm needs to be developed to improve RACH performance.

MULTIPLE DIGITAL CHAINS AT THE BS

One possible way to improve RACH performance (i.e., reduce the required SNR) is to use multiple digital chains at the BS. For example, by steering multiple receive beams in the same direction simultaneously, the received signals can be non-coherently accumulated, resulting in performance gain. In Fig. 5, we compare the performance when the number of receive digital chains at the BS is 1, 2, or 4 while the MS still has a single digital chain. It is seen that about 3 dB gain is obtained in terms of the required SNR satisfying the 1 percent miss detection probability when the number of digital chains is doubled. Since the performance will depend on how the multiple beams are steered, it is an interesting issue to determine the directions of multiple receive beams for the RACH.

EXPLOITING BEAM RECIPROCITY

If the mmWave system operates in TDD mode, one can exploit the channel reciprocity for the random access procedure. If the channel reciprocity holds, the random access procedure can be designed so that the MS transmits a preamble in the best direction only and obtains a high beamforming gain. However, the channel reciprocity may not hold due to different characteristics of the RF circuitry of the transmitter and receiver. Fortunately, for the purpose of a RACH, it will suffice to have beam reciprocity; that is, the best transmit-receive beam pair in downlink is the same as the best receive-transmit beam pair in uplink. Therefore, it is important to calibrate RF circuits for ensuring beam reciprocity, especially in NLOS environments.

CELL DEPLOYMENTS

Since the path loss of an LOS channel is much smaller than that of an NLOS channel, one approach to solve the earlier problems is to carefully deploy BSs in locations where an LOS link between BS and MS can easily be formed. In order to cover a large area with this constraint, however, a large number of BSs will be needed. Therefore, the cell planning method becomes more important in mmWave cellular networks.

Conclusions

An overview of random access in mmWave beamforming cellular networks is presented. We have analyzed the important issues in the design of a random access channel with respect to initial access, handover, uplink-downlink configuration, and scheduling. Through numerical simulations, it is shown that the performance gain from beamforming in multiple directions without knowledge of the best beam pair can still be achieved. As having a large number of antenna elements is not a fundamental solution, we present another approaches and discuss the research challenges. The research on random access for the mmWave cellular network is still at an early stage. We expect that more research on random access will be conducted for successful development of the mmWave cellular network for 5G mobile broadband.

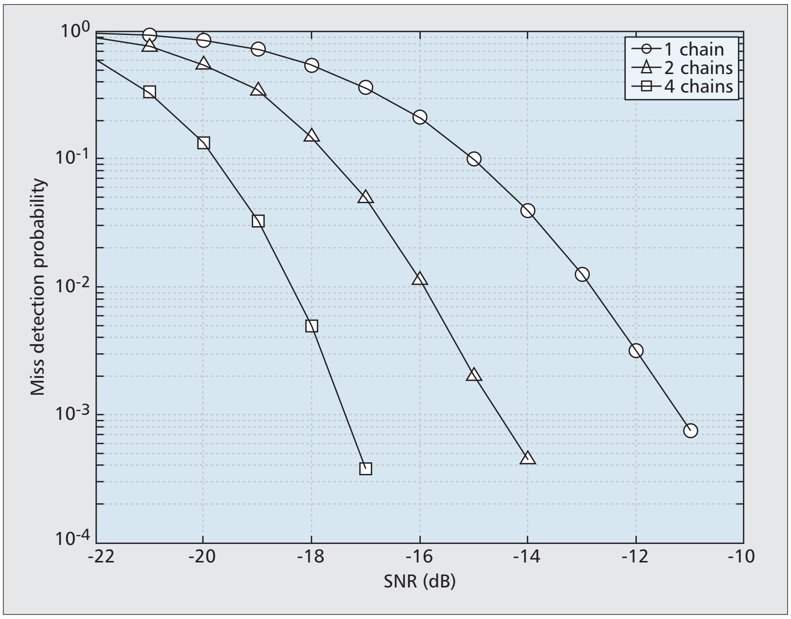


Figure 5. The miss detection probability vs. SNR where $N_{\text{slot}} = 16$. The number of digital chains is 1, 2, or 4.

References

[1] Z. Pi and F. Khan, "An Introduction to Millimeter-Wave Mobile Broadband Systems," *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 101–07.

[2] E. G. Larsson *et al.*, "Massive MIMO for Next Generation Wireless Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 186–95.

[3] 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, pp. 106–13.

[4] T. S. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," *IEEE Access*, vol. 1, May 2013, pp. 335–49.

[5] T. S. Rappaport *et al.*, "Broadband Millimeter-Wave Propagation Measurements and Models Using Adaptive-Beam Antennas for Outdoor Urban Cellular Communications," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, Apr. 2013, pp. 1850–59.

[6] S. Sesia, I. Toufik, and M. Baker, *LTE – The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed., Wiley, 2011.

[7] ITU-R M.2135-1, "Guidelines for Evaluation of Radio Interface Technologies for IMT-Advanced," 2009.

[8] B. Friedlander and S. Scherzer, "Beamforming versus Transmit Diversity in the Downlink of a Cellular Communications System," *IEEE Trans. Vehic. Tech.*, vol. 53, no. 4, July 2004, pp. 1023–34.

[9] D. C. Chu, "Polyphase Codes with Good Periodic Correlation Properties," *IEEE Trans. Info. Theory*, vol. 18, no. 4, July 1972, pp. 531–32.

[10] L. Sanguinetti, M. Morelli, and L. Marchetti, "A Random Access Algorithm for LTE Systems," *Trans. Emerging Tel. Tech.*, vol. 24, no. 1, Jan. 2013, pp. 49–58.

Biographies

Cheol Jeong (cheol.jeong@ieee.org) received his B.S. degree in electrical and electronics engineering from Yonsei University, Seoul, Korea, in 2003, and his Ph.D. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2010. From August 2010 to July 2011, he was with the Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada, as a postdoctoral fellow. In September 2011, he joined the Samsung Electronics, where he is currently a senior engineer. His research interests include MIMO relay communications, physical layer security, and millimeter-wave communications.

Jeongho Park (jeongho.jh.park@samsung.com) received his Ph.D. degree in electronic engineering from Yonsei University. Since he joined Samsung Electronics in 2005, he has mainly been engaged in development of wireless communications and standardization including IMT-Advanced systems. Currently, he is the director of the Communications Research Team at the Samsung Electronics DMC R&D Center, and his research interest includes beyond-4G and 5G technologies.

Hyunkyu Yu received his B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University in 2000, 2002, and 2006, respectively. Since 2006 he has been a senior research engineer at Samsung Electronics. His primary interests are focused on next generation wireless communication systems.

Large-Scale Antenna Systems with Hybrid Analog and Digital Beamforming for Millimeter Wave 5G

Shuangfeng Han, Chih-Lin I, Zhikun Xu, and Corbett Rowell

The authors are with China Mobile Research Institute.

ABSTRACT

With the severe spectrum shortage in conventional cellular bands, large-scale antenna systems in the mmWave bands can potentially help to meet the anticipated demands of mobile traffic in the 5G era. There are many challenging issues, however, regarding the implementation of digital beamforming in large-scale antenna systems: complexity, energy consumption, and cost. In a practical large-scale antenna deployment, hybrid analog and digital beamforming structures can be important alternative choices. In this article, optimal designs of hybrid beamforming structures are investigated, with the focus on an *N* (the number of transceivers) by *M* (the number of active antennas per transceiver) hybrid beamforming structure. Optimal analog and digital beamforming designs in a multi-user beamforming scenario are discussed. Also, the energy efficiency and spectrum efficiency of the $N \times$ M beamforming structure are analyzed, including their relationship at the green point (i.e., the point with the highest energy efficiency) on the energy efficiency-spectrum efficiency curve, the impact of *N* on the energy efficiency performance at a given spectrum efficiency value, and the impact of *N* on the green point energy efficiency. These results can be conveniently utilized to guide practical LSAS design for optimal energy/spectrum efficiency trade-off. Finally, a reference signal design for the hybrid beamform structure is presented, which achieves better channel estimation performance than the method solely based on analog beamforming. It is expected that large-scale antenna systems with hybrid beamforming structures in the mmWave band can play an important role in 5G.

INTRODUCTION

Studies on the fifth generation (5G) wireless communication system are gaining more momentum worldwide in an attempt to provide solutions for the exponential increase of mobile data traffic by 2020. Although 5G is in its embryonic stage, some trends have appeared in how to design 5G networks, such as to make it green and soft, as proposed by China Mobile [1]. Multiple research topics have been identified as 5G candidates, including large-scale antenna systems (LSAS), non-orthogonal multiplex access, full duplex, spectrum sharing, high-frequency bands (e.g., millimeter-wave, mmWave), high-density networks, new network architecture, new waveform design, and so on. These technologies may have great potential in system performance improvement in 5G.

The fundamental premise of LSAS [2] is that the number of base station antennas is much larger than the number of single antenna terminals. Theoretically, LSAS with full digital beamforming (BF) can yield the optimal performance, significantly increasing the system energy efficiency (EE) and spectrum efficiency (SE) via multiuser BF [3]. However, implementing LSAS in lower frequency bands is difficult since the much larger physical footprints of LSAS base stations will not only bring significant tower construction challenges, but also lead to increasing concerns about possible health effects. Higher frequency bands like mmWave [4] are appealing for LSAS, since the physical array size of LSAS can be greatly reduced due to the decrease in wavelength. Another differentiating feature that makes mmWave especially attractive in 5G is that large chunks of underutilized spectrum are available, which can be potentially utilized to provide significant system capacity improvements. Recently, mmWave outdoor channel measurement results [5], advances in hardware design [6], a successful outdoor trial [7], and the availability of abundant spectrum in mmWave have encouraged the wireless industry to consider mmWave for cellular systems [8].

When a large number of antennas are implemented to achieve better BF gains, however, implementing the same number of transceivers may not be feasible due to excessive demand on real-time signal processing for high BF gains [11], high power consumption, and high cost (especially the high cost and power consumption of mixed-signal devices in mmWave systems). A BF structure with a much lower number of digital transceivers than the total antenna number will therefore be more practical and cost effective to deploy. One interesting approach to reducing the transceiver number is via analog BF [6, 7], where each transceiver is connected with multiple active antennas, and the signal

0163-6804/15/$25.00 © 2015 IEEE

phase on each antenna is controlled via a network of analog phase shifters. Generally, each transceiver generates one beam toward one user in analog BF. When the user number simultaneously served is much smaller than the antenna number (this is generally true in an LSAS system), the transceiver number can be designed to be much smaller than the antenna number. However, there may be severe inter-user interference when users are not adequately spatially separated. Digital BF over transceivers can then be utilized to achieve multiple data stream precoding on top of analog BF [9–12] to further enhance the performance.

Two hybrid BF structures that have drawn much attention of researchers are shown in Fig. 1, with N being the transceiver number and NM being the antenna number. In structure 1 each transceiver is connected with all antennas, such that the transmitted signal on each of the N digital transceivers goes through NM RF paths (mixer, power amplifier, phase shifter, etc.) and summed up before being connected with each antenna element [9, 10], as shown in Fig. 1a. Analog BF is performed over NM RF paths per transceiver, and digital BF can then be performed over N transceivers. This structure is a natural combination of analog BF and digital BF, and achieves full BF gain for each transceiver. However, the complexity of this structure is rather high; for example, the total number of RF paths is N^2M.

An $N \times M$ hybrid BF structure is shown in Fig. 1b, where each of the N transceivers is connected to M antennas. Analog BF is performed over only M RF paths in each transceiver, and digital BF is performed over N transceivers [8, 11]. This structure is more practical for base station antenna deployment in the current cellular systems, where each transceiver is generally connected to a column of antennas. Compared to structure 1, the BF gain per transceiver is $1/N$ the gain in structure 1, but with much reduced complexity, the total number of RF paths being NM.

Recently, there has been growing interest in hybrid BF structure design for mmWave communication. For structure 1, a simple precoding solution was proposed assuming only partial channel knowledge at the base station and mobile station in the form of angle of arrival (AoA) and angle of departure (AoD) knowledge [9]. The spatial structure of mmWave channels was further exploited in [10] to formulate the single-user precoding/combining problem as a sparse reconstruction problem. AoA estimation and beamforming algorithms were proposed in [11] for structure 2. A successful mmWave outdoor trial was carried out in Korea, in which the hybrid BF of structure 2 was implemented [7]. There are still many challenging issues regarding hybrid BF structures:

- What is the performance gap compared to digital BF?
- Does a larger transceiver number N always lead to better EE, and what is the EE-SE optimal design?
- How can we design efficient reference signals (RSs) for better availability of the channel state information at the transmitter side?

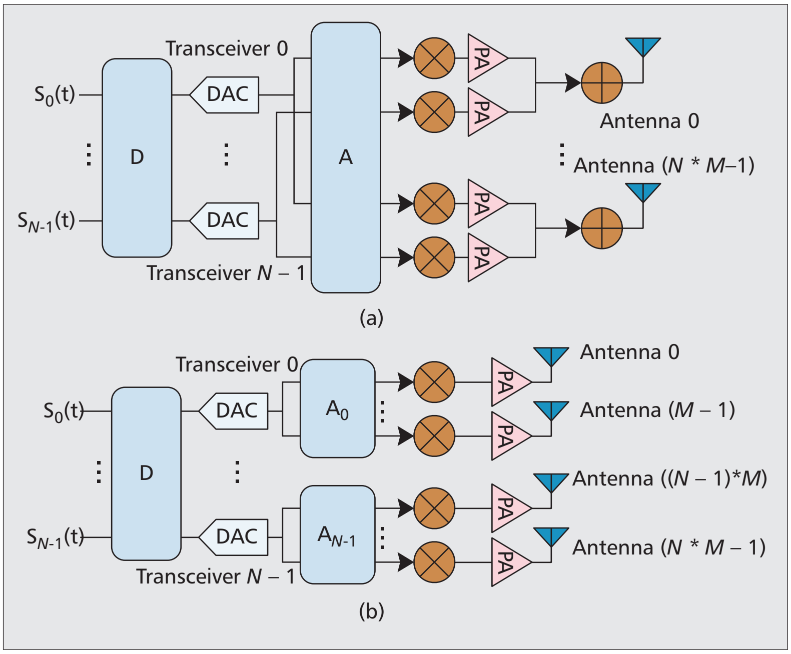


Figure 1. Hybrid BF structures.

This article addresses the above important issues on how LSAS with hybrid BF structure can potentially be utilized in mmWave systems with a focus on the N by M hybrid BF structure. The optimal analog BF and digital BF design are investigated for a multi-user BF scenario. The EE-SE relationship of the $N \times M$ hybrid BF structure is analyzed, paving a path to an EE-SE optimized design. A downlink RS design is discussed. Beam domain RSs based on hybrid BF are presented, which outperform RSs based on analog BF. This article is finally summarized.

Hybrid Analog BF and Digital BF Design

This section studies the hybrid BF strategies for the two structures shown in Fig. 1, especially for the case of a sufficiently large antenna number. Suppose there are N downlink users, each having a single antenna. Denote $\mathbf{H}$ as the downlink channel with the size of $N \times NM$, $\mathbf{A}$ as the analog BF matrix with the size of $NM \times N$, $\mathbf{D}$ as the digital BF matrix with the size of $N \times N$, $\mathbf{s}$ as the N user data vector with the size of $N \times 1$, and $\mathbf{n}$ as the noise vector, respectively. The received signal in the downlink is expressed as $\mathbf{y} = \mathbf{HADs} + \mathbf{n}$.

Design for Structure 1

As shown in Fig. 1a, the digital signal from each transceiver can be delivered to any antenna after analog weighting. The ith column of $\mathbf{A}$ is thus the analog BF vector for all NM antennas on the ith transceiver, $i = 0, \ldots, N - 1$. From matrix theory, it can be derived that the channel capacity can be achieved when the hybrid BF transmits signals over the largest N eigenvectors of $\mathbf{H}$. The power allocation can be carried out following the water-filling method. A good reference on joint $\mathbf{D}$ and $\mathbf{A}$ design can be found in [10].

The analytical goal of the first case is to find the optimal number of transceivers and, correspondingly, the optimal number of antennas per transceiver given the total number of antennas, while the investigation in the second case is to explore how the independent N *and* M *should be jointly optimized.*

Note that the upper bound on the channel capacity with effective channel **HA** can be achieved when **HA** has N equal singular values [2]. When M is large enough, **HA** turns out to be a diagonal matrix with equal diagonal elements if we take **A** to be exactly the normalized conjugate transpose of **H**. By doing so, there is negligible inter-user interference, and hence the upper bound of N-user capacity can be achieved. Correspondingly, the optimal **D** is an identity matrix, indicating that there is actually no need to implement digital BF across the transceivers. Although the optimal N user channel capacity can be achieved under this structure, the analog weighting should be done N^2M times in total, which brings high implementation complexity.

Design for Structure 2

For the $N \times M$ hybrid BF structure in Fig. 1b, the digital signal from each transceiver can only be delivered to M antennas. Therefore, analog BF matrix **A** has a different structure from that in structure 1, $\mathbf{A} = diag[\mathbf{A}_0, \ldots, \mathbf{A}_{N-1}]$, where $\mathbf{A}_i$, $i = 0, \ldots, N-1$ is the $M \times 1$ analog precoder on the ith transceiver. Suppose **H** and **A** are known; the digital BF matrix **D** can then be designed with traditional multiple-input multiple-output (MIMO) theories (e.g., to maximize the sum rate). To the best of the authors' knowledge, the joint optimal design of **D** and **A** is still an open issue.

When M is large enough, by partitioning the ith row of **H** (denoted by $\mathbf{h}_i$) into N consecutive parts with equal number of elements, and designing $\mathbf{A}_i$ to be exactly the normalized conjugate transpose of the ith part of $\mathbf{h}_i$, again **HA** becomes exactly a diagonal matrix with equal diagonal elements. The capacity upper bound can then be achieved. Correspondingly, the optimal **D** is also the identity matrix. Note that the diagonal elements of **HA** are actually the BF gains, which is N times higher in structure 1 than that in structure 2. This leads to roughly $Nlog_2N$ b/s/Hz capacity loss in structure 2.

EE-SE Analysis of the $N \times M$ BF Structure

From earlier, when the user number equals the transceiver number, structure 1 can achieve the same channel capacity as that achieved by the traditional digital BF structure; whereas the capacity achieved by structure 2 is lower than that achieved in structure 1. For both structures, the required number of transceivers is much smaller than that in the digital BF structure. This section continues to investigate the optimal transceiver number N in terms of EE-SE co-design [1]. The analysis is focused on structure 2 (i.e., the $N \times M$ BF structure) since its complexity is much lower than that of structure 1, and the capacity loss is not significant with a practical N.

EE-SE Relationship

Consider the $N \times M$ BF structure: perfect analog BF is assumed within M antennas per transceiver for one user (in total there are N users). There is zero inter-user interference with the assumption of a large enough M or via proper user scheduling even when M is not so large. For example, the beamwidth of the analog beam generated by a 50-element linear antenna array with half wavelength spacing can be as small as 2°. It would not be difficult to schedule, say, $N = 10$ users with negligible inter-user interference in one cell. The N user sum capacity of this structure is derived as $C = WNlog_2(1 + MP\eta_{SE}/WN_0)$, where W is the system bandwidth, P is the total power of M power amplifiers (PAs) per transceiver, η_{PA} is the PA efficiency, and N_0 is the thermal noise density. Channel gain is assumed to be 1. The SE of this structure is written as $\eta_{SE} = C/W$.

An accurate power model is needed to calculate the EE. However, this is not straightforward, since base stations have different types (macro, pico, femto) and are generally produced by different vendors with various implementation technologies. In this section, the following simple power model is used — $P_{total} = NP + P_{static} = NP + NP_0 + P_{common} + NMP_{rf\_circuit}$ — where P_{total} is the total power, NP is the RF power of total N transceivers, P_{static} is the static power of the base station, including the part of power NP_0 that scales with the number of transceivers N, P_{common}, which is common for any transceiver number, and $NMP_{rf\_circuit}$, which scales with total antenna number NM. The EE-SE relationship can be written as

$$\eta_{EE} = C / P_{total} = \frac{\eta_{SE}}{\left(2^{\frac{\eta_{SE}}{N}} - 1\right)\frac{N_0}{\eta_{PA}}\frac{N}{M} + \frac{NP_0 + P_{common} + NMP_{rf\_circuit}}{W}} \quad (1)$$

EE-SE Relationship at the Green Points

As shown in [13], the EE-SE relationship based on Shannon's theory is monotonic, where a higher SE will always lead to a lower EE. When the circuit power is considered, however, there is a green point on the EE-SE curve where the maximum EE, η^*_{EE}, is achieved. Two cases are discussed here for the $N \times M$ hybrid BF structure, the case when $NM = L$ (i.e. the total antenna number is fixed to be L, but N and M are variable), and the case when N and M are independent. The analytical goal of the first case is to find the optimal number of transceivers and, correspondingly, the optimal number of antennas per transceiver given the total number of antennas, while the investigation on the second case is to explore how the independent N and M should be jointly optimized.

It can be derived [12] based on Eq. 1 that there is only one green point on the EE-SE curve for each case (i.e., there is only one η^*_{SE} that maximizes the EE performance). The relationship between η^*_{EE} and η^*_{SE} is further given as $lg(\eta^*_{EE}) = -\eta^*_{SE}\, lg2/N + lg(M\eta_{PA} /N_0 ln2)$; that is, $lg(\eta^*_{EE})$ scales with η^*_{SE} linearly with a slope of $-lg2/N$. Similar to the EE-SE relationship with classic Shannon theory, a higher η^*_{SE} will always lead to a lower η^*_{EE}. Interestingly, the relationship between η^*_{EE} and η^*_{SE} is independent of P_0, P_{common}, $P_{rf\_circuit}$, and W, although, as can be seen from Eq. 1, η^*_{SE} and η^*_{EE} are determined

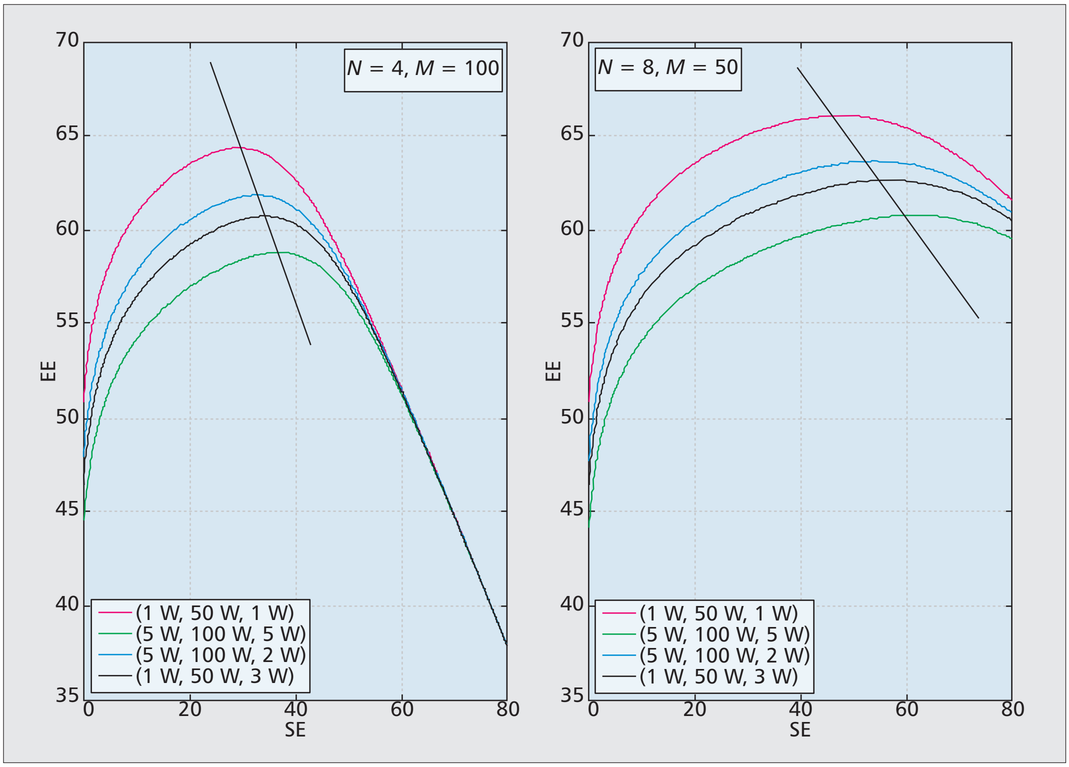


Figure 2. EE-SE relationship at the green points.

This indicates that in the EE optimal design, a smaller transceiver number *N* brings more EE performance improvement with a given SE reduction.

based on all the other parameters. The EE-SE relationship was also investigated in [14] for MIMO systems, which also showed a similar trend.

Assume $W = 2 \times 10^7$ Hz, $N_0 = 10^{-17}$ dBm/Hz, and a channel gain of –100 dB. The EE-SE relationship is depicted for two scenarios, the scenario with $N = 4$ and $M = 100$, and the scenario with $N = 8$ and $M = 50$. For each scenario, four ($P_{rf\_circuit}$, P_{common}, P_0) combinations are simulated, including (1 W, 50 W, 1 W), (5 W, 100 W, 5 W), (5 W, 100 W, 2 W), and (1 W, 50 W, 3 W). EE (bits per Joule) is depicted in log-scale, and SE (bits per second per Hertz) is depicted in linear scale. As shown in Fig. 2, the EE-SE curve is different with different ($P_{rf\_circuit}$, P_{common}, P_0) combinations. On each curve, there is only one green point. Also, the green points are actually in a straight line for both scenarios, with a large slope in the first scenario. This indicates that in the EE optimal design, a smaller transceiver number *N* brings more EE performance improvement with a given SE reduction.

With the above analysis, therefore, it is expected that the system operates at the green point. Also, it is important that η^*_{SE} satisfies the system SE requirement. Besides, η^*_{EE} should be high enough. These require careful design of the following parameters: P_0, P_{common}, $P_{rf\_circuit}$, W, η_{PA}, N, and M. For example, when other parameters are given, N can be designed to maximize η^*_{EE}. The optimal *N* in terms of EE-SE co-design is discussed in the following.

How Does *N* Affect EE-SE?

When the required SE is predetermined, it is desirable that the transceiver number *N* is optimized, yielding the highest EE performance with the minimum transceiver number. Based on Eq. 1, it is found that in the cases $NM = L$ and independent *N* and *M*, for any given SE, there is only one optimal *N* to yield the best EE. A detailed proof can be found in [12]. The practical meaning of the existence of the optimal *N* is that with a given SE, a system designer does not need to implement too many transceivers to achieve the best EE performance.

Assume $P_{rf\_circuit} = 1$ W, $P_{common} = 50$ W, $P_0 = 1$ W, $\eta_{PA} = 0.375$, $W = 2 \times 10^7$ Hz, $N_0 = 10^{-17}$ dBm/Hz, and a channel gain of –100 dB. Considering the $NM = 500$ case, the impact of *N* (from 1 to 10) on EE performance is shown in subplot 1 of Fig. 3, where five SE values are simulated. Note that since *N* and *M* are integers, there are only five valid (*N*, *M*) combinations for $NM = 500$, that is, (1,500), (2,250), (4,125), (5,100), and (10,50). It can be observed that on each curve there is one optimal *N* that yields the highest EE. For example, when SE is 20 b/s/Hz, the optimal *N* is 4. When SE is 8 b/s/Hz, the optimal *N* is just 1. The case when *N* is larger than 10 is not shown in the figure, since it may be difficult to schedule users with negligible inter-user interference when *M* is very small.

When *N* and *M* are independent, the impact of *N* on EE performance is shown in subplot 2 in Fig. 3, where $M = 50$, and other parameters are the same as those in subplot 1. Similar to the fixed *NM* case, on each curve there is one optimal *N*. For example, when SE is 40 b/s/Hz, the optimal *N* is 6. Different from the fixed *NM* case, the EE performance is very sensitive to *N*, because the total antenna number scales with *N*.

In a practical system operation, the SE requirement may vary according to the traffic load and service types. For example, as shown in

The intuition is that given one required SE value, we need to optimize the system parameters such that the EE achieved via the optimal transceiver number N is exactly at the green point of its EE-SE curve.

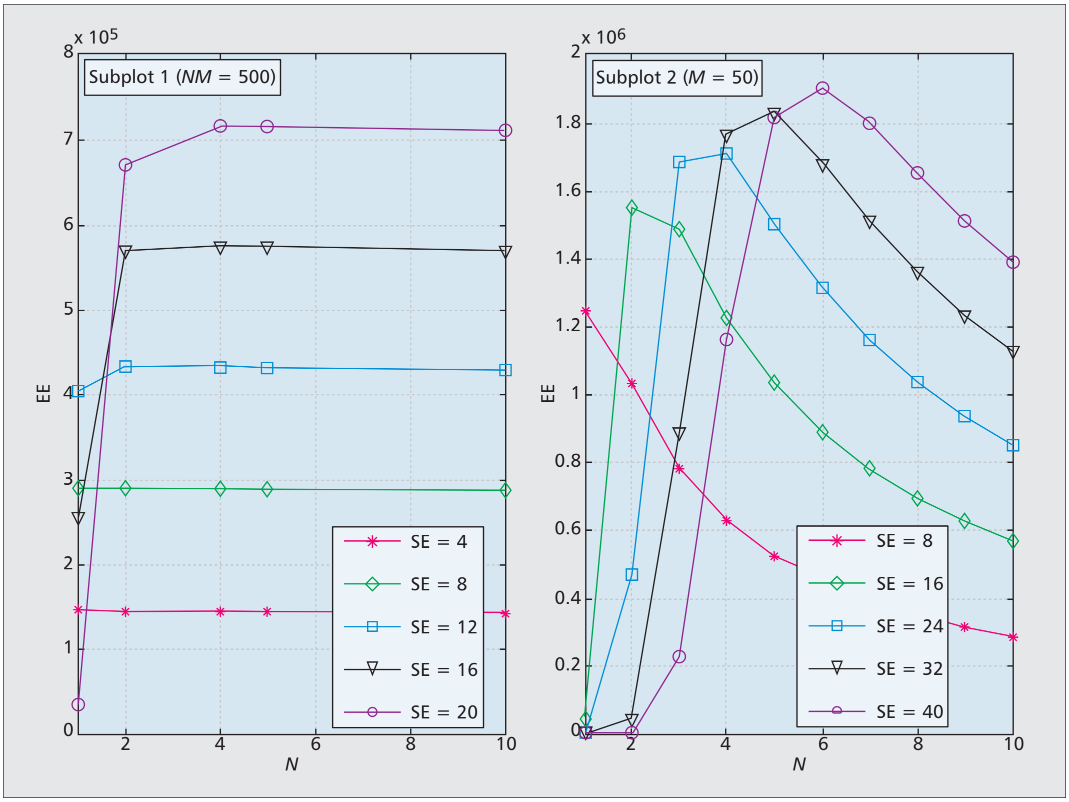


Figure 3. N vs. EE with different SE values.

subplot 2 in Fig. 3, with the maximum 40 b/s/Hz SE requirement, the optimum N should be designed to be 6. But when the SE requirement is reduced to 8 b/s/Hz, the optimal N should be 1. Therefore, it is important that for the possible SE range, the system can be designed with the largest optimal N, and selects the best N according to the SE requirement via transceiver on/off. This can help to further enhance the EE performance according to the system traffic load.

Relationship between the Green Point EE and N

Given the transceiver number N, there is only one green point on EE-SE curve. We have shown that for any SE, there is one optimal N. But it is still unclear whether $\eta^*_{EE}(N)$ at the green point is the highest EE performance of all N. Based on Eq. 1, it can be derived that when N and M are independent, a larger N always results in a higher η^*_{EE}. When $NM = L$, however, the impact of N on $\eta^*_{EE}(N)$ is determined based on the parameters in Eq. 1 [12].

One example is shown to illustrate how N impacts $\eta^*_{EE}(N)$. Assume $P_{rf\_circuit} = 1$ W, $P_{common} = 50$ W, $P_0 = 1$ W, $\eta_{PA} = 0.375$, $W = 2 \times$ 108 Hz, $N_0 = 10^{-17}$ dBm/Hz, and a channel gain of –100 dB. Note that a much larger bandwidth is considered here for a smaller optimal N in the simulation. The EE-SE curves with different N are shown in subplot 1 of Fig. 4 for $NM = 800$, with N being 1, 2, 4, 5, 8, 10, and 16. It can be found that as N increases from 1 to 8, $\eta^*_{EE}(N)$ increases, but when N increases from 8 to 16, $\eta^*_{EE}(N)$ decreases. The optimal transceiver number is therefore 8. Transceiver numbers larger than 16 are not shown in this figure. For the independent N and M case, the impact of N (from 1 to 10) on $\eta^*_{EE}(N)$ is shown in subplot 2, with $M = 50$. As N increases, the green point EE also increases monotonically. The intuition is that given one required SE value, we need to optimize the system parameters such that the EE achieved via the optimal transceiver number N is exactly at the green point of its EE-SE curve.

Reference Signal Design

Based on the earlier analysis, hybrid BF structures can achieve optimal channel capacity with a much smaller number of transceivers N optimized in terms of joint EE-SE co-design. This section discusses possible reference signal design, which determines to what extent the LSAS with hybrid BF structure helps to enhance system performance.

Generally, there are two methods to obtain downlink CSI: via uplink sounding in a time-division duplex (TDD) system or downlink RSs. The first method is not trivial in LSAS, because the downlink/uplink channel reciprocity calibration using traditional methods at a base station may not function well due to high complexity of the calibration circuits. In addition, the $N \times M$ structure makes it difficult to accurately measure the CSI per antenna. The second method has issues as well, since the antennas connected to each transceiver are actually one logical antenna, and how the reference signals should be transmitted is still not well understood. Therefore, one prac-

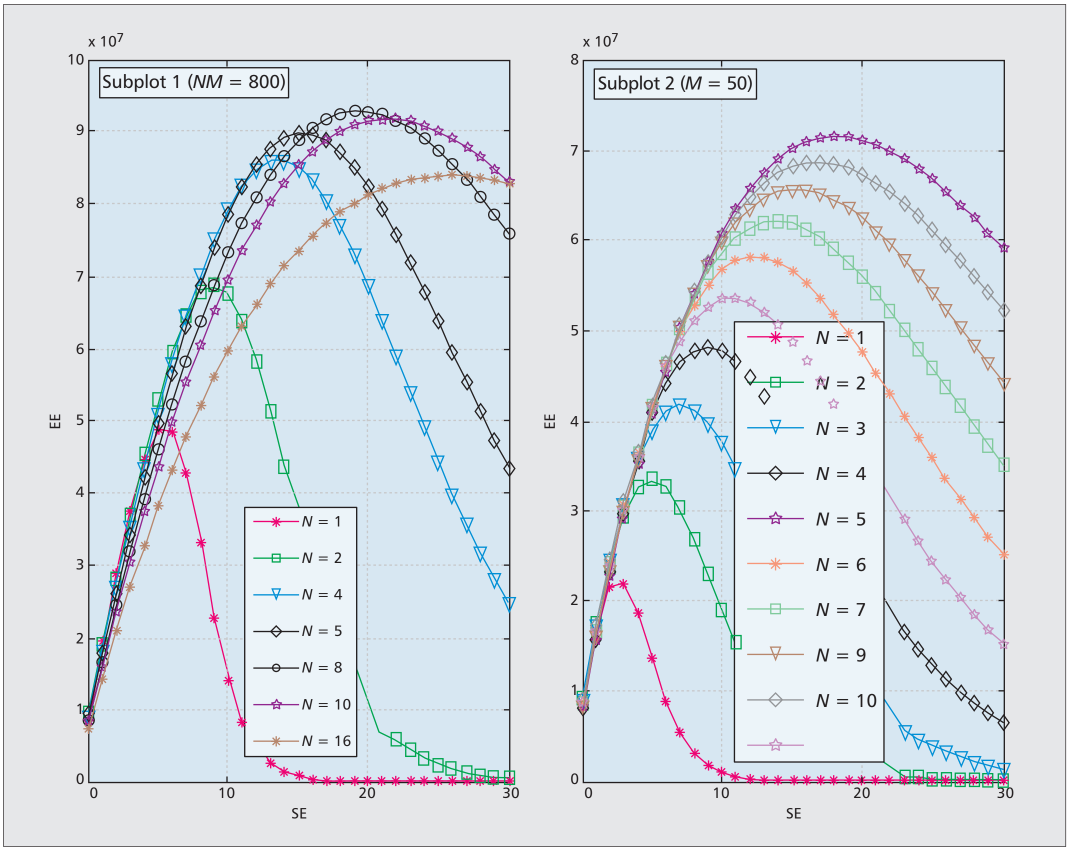


Figure 4. EE-SE curves with different N.

The accuracy of this design depends on the RSs coverage and the transceiver number N. With a given transceiver number N, it is difficult to improve AoD estimation accuracy. In the next subsection, one beam domain RSs design with hybrid BF is presented.

tical assumption on channel information availability in LSAS with hybrid BF structure is that partial channel information like AoA and AoD are known, as used in [9]. Two beam domain downlink RS designs for the AoD estimation, RSs with analog BF and RSs with hybrid BF, are discussed below.

Beam Domain RSs with Analog BF

Since per antenna channel state information (CSI) is difficult to measure accurately, beam domain RS is more practical for this $N \times M$ BF structure, especially in line of sight (LOS) environments. One straightforward design is that each transceiver transmits one predefined beam via analog BF over M antennas, such that there are N simultaneous beams with different main beam directions (i.e., AoD in the downlink). Mobile users then feed back the index of the received beam with the highest signal power, and the base station will transmit to each user with the corresponding beam. The accuracy of this design depends on the RSs' coverage and transceiver number N. With a given transceiver number N, it is difficult to improve AoD estimation accuracy. In the next subsection, one beam domain RSs design with hybrid BF is presented.

Beam Domain RSs with Hybrid BF

Consider the $N \times M$ structure with a linear antenna array in an orthogonal frequency-division multiplex (OFDM) system. As shown in Fig. 5a, on the kth subcarrier the reference signal s_k first passes through the digital BF, and then, after the digital-to-analog conversion, passes through the analog BF. The same analog BF with a main beam direction ϕ_0 is applied to all transceivers, while the digital BF matrix $\mathbf{D}_k$ is diagonal with the ith ($i = 1, \ldots, N - 1$) diagonal element being $\exp(j\alpha_i)$, where α_i is the phase shift on the ith transceiver. By choosing proper phase shift on each transceiver, the sum array factor (AF) of all NM antennas in a direction ϕ around ϕ_0 can be maximized, with the corresponding maximum value being N times that of a single transceiver analog BF. More details on hybrid BF design can be found in [15] for both the linear array and planar array cases. Therefore, for a given analog BF, the main beam direction of the hybrid BF is determined by the digital BF weights. This leads to the following beam domain RS design based on hybrid BF.

As shown in Fig. 5b, the RSs occupy $2K + 1$ consecutive subcarriers, which are assumed to be well within the channel's coherent bandwidth. In the sth OFDM symbol, the main beam direction of the analog BF is set to be $\phi_{0,s}$ for all transceivers. While the main beam direction of the hybrid analog and digital BF on the kth ($k = 0, 1, \ldots 2K$) subcarrier is $\phi_{0,s} + (K - k)\,\Delta$, with Δ being the beam spacing. The digital BF weight on the ith transceiver on the kth subcarrier in the sth OFDM symbol can readily be calculated accordingly. With the feedback of the subcarrier index on which the signal power is the largest, the base station knows the AoD information, such as the kth subcarrier corresponds to an AoD of $\phi_{0,s} + (K - k)\,\Delta$. This is because within the coherent bandwidth the channel response is the same for all subcarriers, and therefore, the

The design based on hybrid BF is featured by same analog BF on each transceiver, on top of which digital BF is designed to maximize the gain in a certain direction around the main beam direction of the analog BF. Much better AoD estimation performance can be achieved compared to beam domain RSs with analog BF.

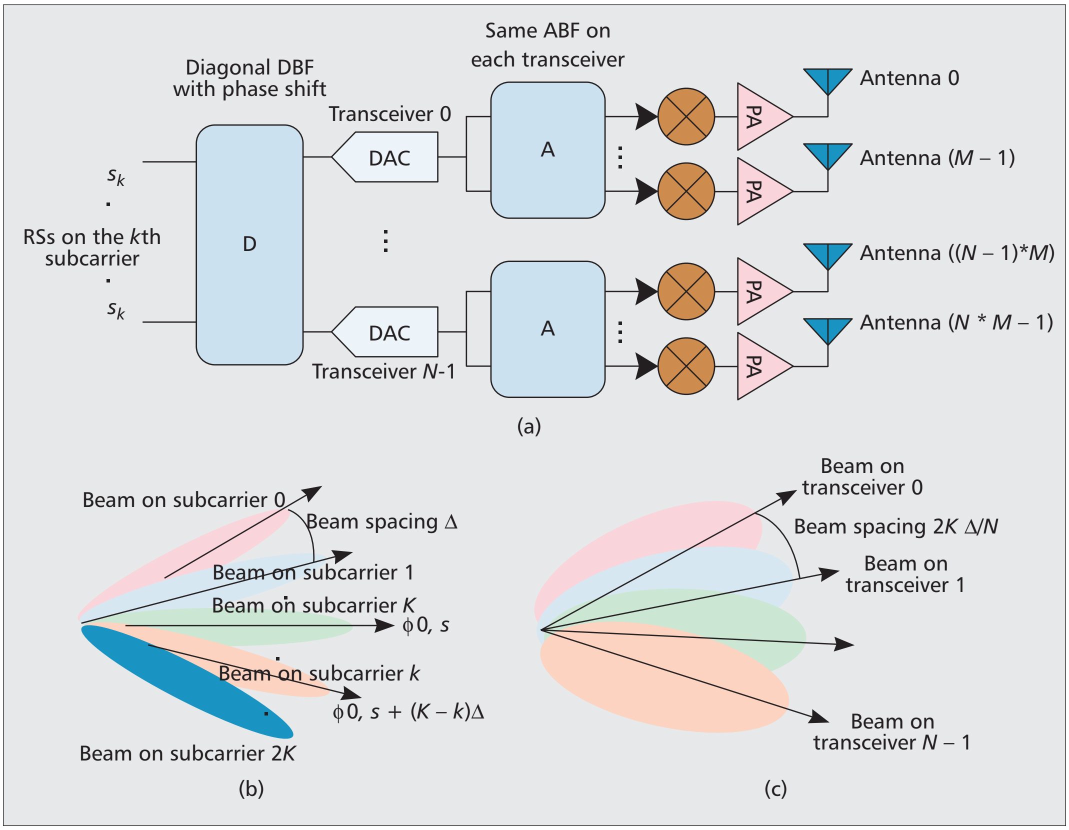


Figure 5. RS design for the $N \times M$ hybrid BF structure: a) hybrid BF for beam domain RSs; b) beam domain RSs with hybrid BF on different subcarriers; c) beam domain RSs with M antenna analog BF.

difference in received signal power on different subcarriers only comes from different digital BF weights.

AoD Estimation Performance Comparison

Consider the cell area covered by $2K + 1$ hybrid analog and digital BF beams shown in Fig. 5b, which is roughly $2K\Delta$. The AoD estimation accuracy of RSs with hybrid BF is therefore Δ. As shown in Fig. 5c, however, for RS beams with analog BF per transceiver (totally N beams) to cover the same area, the AoD estimation accuracy is $2K\Delta/N$. In LOS environments where the channel response is rather flat in frequency domain, $2K$ can be very large, indicating that RSs design with hybrid BF has much better AoD estimation performance than RSs with analog BF. According to the measurement results on 38 GHz [5], more than 80 percent of the non-LOS (NLOS) links had root mean square (RMS) delay spreads under 20 ns, and 90 percent of the NLOS links had RMS delay spreads under 40 ns. This indicates that the coherent bandwidth in NLOS environments can also be very large. Therefore, a large $2K/N$ is possible even in NLOS environments.

Illustration of Hybrid BF

The effect of hybrid BF is shown in Fig. 6. A 32-element linear antenna array (with half wavelength antenna spacing) is simulated, where the transceiver number is 4, and antenna number per transceiver is 8. The max array gain is achieved at azimuth 90°. The first curve is AF of 8 antenna elements analog BF with main beam direction at azimuth 90°. The second curve is the AF envelope of the 32-element hybrid analog and digital BF. It can be seen that the amplitude of the second curve is exactly four times that of the first curve. With same analog BF per transceiver, seven digital BF designs are shown, with main beam direction of azimuth 84°, 86°, 88°, 90°, 92°, 94, and 96°, respectively. As shown in Fig. 6, the main beam direction can be controlled with hybrid BF design.

Some observations are summarized:

- The further the main beam direction of the hybrid BF is away from the analog BF main beam direction, the smaller the gain. Therefore, the coverage of the beams is limited.
- The hybrid BF designed for 90° is actually the analog BF with all antennas.
- Users located at certain AoD ϕ within the range of 84° to 96° will find the hybrid BF designed for direction ϕ has the largest signal power.

Conclusions

The marriage of LSAS and mmWave technologies is expected to bring significant EE and SE enhancement to 5G. The high cost and power consumption of mixed-signal devices in mmWave systems, however, makes hybrid BF structure with a much reduced transceiver number a possible solution. In an attempt to shed some insight on how hybrid BF can be potentially utilized in mmWave 5G, this article has addressed some important issues. We have discussed the optimal digital and analog BF design, showing that the investigated hybrid BF structures can achieve the multi-user channel capacity. The EE-SE relationship of the $N \times M$ hybrid BF structure have been analyzed, paving a path for an EE-SE optimized design. Finally, we have discussed beam domain RSs. The design based on hybrid BF is featured by the same analog BF on each transceiver, on top of which digital BF is designed to maximize the gain in a certain direction around the main beam direction of the analog BF. Much better AoD estimation performance can be achieved compared to beam domain RSs with analog BF.

Future work may include more efficient single-user and multi-user BF schemes, RS design and feedback mechanism, frame structure, protocol, and signaling design. Also, more effort is needed to investigate the impact of system parameters like W, P_0, $P_{rf\_circuit}$, P_{common}, and antenna spacing on optimal N and M design in terms of joint EE-SE optimization. It is expected that LSAS with mmWave can play an important role in future cellular communication systems.

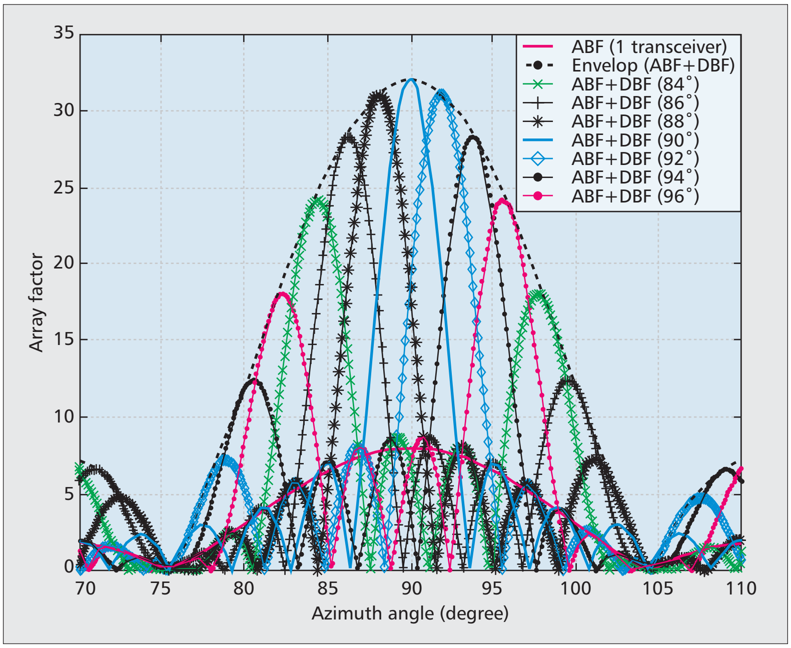


Figure 6. Hybrid BF with a 32-element antenna array.

Acknowledgment

The authors would like to thank the editors and reviewers for their very helpful comments and reviews. Their suggestions greatly improved the presentation of this article. The authors also would like to express their gratitude to Profs. Jing Wang and Shidong Zhou from Tsinghua University for their enlightening comments. The authors are also grateful to the team members in the Green Communication Research Center of the China Mobile Research Institute, particularly Drs. Sen Wang and Zhengang Pan.

References

[1] C.-L. I *et al.*, "Toward Green & Soft: A 5G Perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 66–73.

[2] F. Rusek *et al.*, "Scaling up Mimo: Opportunities and Challenges with very Large Arrays," *IEEE Sig. Proc. Mag.*, vol. 30, no. 1, Jan 2013, pp. 40–60.

[3] H. Q. Ngo, E. G. Larsson, and T. Marzetta, "Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems," *IEEE Trans. Commun.*, vol. 61, no. 4, Apr. 2013, pp. 1436–49.

[4] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges," *Proc. IEEE*, vol. 102, no. 3, 2014, pp. 366–85.

[5] T. Rappaport *et al.*, "Broadband Millimeter-wave Propagation Measurements and Models Using Adaptive-beam Antennas for Outdoor Urban Cellular Communications," *IEEE Trans. Antennas and Propagation*, vol. 61, no. 4, 2013, pp. 1850–59.

[6] C. Doan *et al.*, "Design Considerations for 60 GHz CMOS Radios," *IEEE Commun. Mag.*, vol. 42, no. 12, 2004, pp. 132–40.

[7] 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, pp. 106–13.

[8] Z. Pi and F. Khan, "An Introduction to Millimeter-Wave Mobile Broadband Systems," *IEEE Commun. Mag.*, vol. 49, no. 6, 2011, pp. 101–07.

[9] A. Alkhateeb *et al.*, "Hybrid Precoding for Millimeter Wave Cellular Systems with Partial Channel Knowledge," *Info. Theory and Applications Wksp.*, 2013.

[10] O. El Ayach *et al.*, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, Mar. 2014, pp.1499–1513.

[11] X. Huang, Y. Jay Guo, and J. Bunton, "A Hybrid Adaptive Antenna Array," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, May 2010, pp.1770–79.

[12] S. Han *et al.*, "Large Scale Antenna System with Hybrid Digital and Analog Beamforming Structure," *ICC Wksp. 2014*.

[13] G. Y. Li*et al.*, "Energy-Efficient Wireless Communications: Tutorial, Survey, and Open Issues," *IEEE Wireless Commun.*, vol. 18, no. 6, Dec. 2011, pp. 28–35.

[14] Z. Xu, Z. Pan, and C.-L. I, "Fundamental Properties of the EE-SE Relationship," *IEEE WCNC 2014*.

[15] S. Han *et al.*, "Reference Signals Design for Hybrid Analog and Digital Beamforming," *IEEE Commun. Letters*, vol. 18, no. 7, July 2014, pp. 1191–93.

Biographies

Shuangfeng Han (hanshuangfeng@chinamobile.com) is currently a senior project manager in the Green Communication Research Center of the China Mobile Research Institute (CMRI). His research interests are mainly focused on green technologies R&D in 5G wireless communication systems, including large-scale antenna systems, active antenna systems, co-frequency co-time full duplex, non-orthogonal multiple access schemes, energy efficiency and spectrum efficiency co-design, and beyond cellular green generation solutions. Prior to joining CMRI, he was a senior engineer at Samsung Electronics' Research Center from 2006 to 2012. His research interests include MultiBS MIMO, MIMO codebook design, small cell/HetNet, millimeter-wave communication, D2D, and distributed radio over fiber. He graduated from Tsinghua University, Beijing, in 2006, and majored in information and communication systems. He is an inventor on 37 patent applications, and the author on over 20 peer-reviewed conference and journal publications. He has been reviewer for various IEEE journals and conferences.

Chih-Lin I (icl@chinamobile.com) is the chief scientist of China Mobile Wireless Technologies, in charge of advanced wireless communication R&D efforts of CMRI. She established the Green Communications Research Center of China Mobile, spearheading major initiatives including 5G Key Technologies R&D; high energy efficiency system architecture, technologies, and devices; green energy; C-RAN and soft base station. She received her Ph.D. degree in electrical engineering from Stanford University, and has almost 30 years of experience in wireless communications. She has worked in various world-class companies and research institutes, including the wireless communication fundamental research department of AT&T Bell Labs; the headquarters of AT&T as director of Wireless Communications Infrastructure and Access Technology; ITRI of Taiwan as director of Wireless Communication Technology; and Hong Kong ASTRI as VP and the founding general director of the Communications Technology Domain. She received the *IEEE*

Transactions on Communications Stephen Rice Best Paper Award, and is a winner of the CCCP National 1000 Talent program. She was an elected Board Member of IEEE ComSoc, Chair of ComSoc Meetings and Conference Board, and Founding Chair of the IEEE WCNC Steering Committee. She is currently Chair of FuTURE Forum 5G SIG, an Executive Board Member of GreenTouch, a Network Operator Council Member of ETSI NFV, a Steering Board Member of Wireless World Research Forum, and an adjunct professor of Beijing University of Posts and Telecommunications.

ZHIKUN XU (xuzhikun@chinamobile.com) received his B.S.E. and Ph.D. degrees in signal and information processing from Beihang University (BUAA), Beijing, 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 (GCRC) of CMRI as a project manager. His current interests include green technologies, the fundamental relationships between energy efficiency and spectral efficiency, energy-efficient network deployment and operation, cross-layer resource allocation in cellular networks, and advanced signal processing and transmission techniques.

CORBETT ROWELL (corbett.rowell@nu.edu.kz, corbettrowell@chinamobile.com) is a professor of electronic and electrical engineering at Nazarbayev University. He received his B.A. degree (honors) in physics from the University of California Santa Cruz, and his M.Phil. and Ph.D. degrees in electrical and electronic engineering from Hong Kong University of Science and Technology and Hong Kong University, respectively. He has worked for over 18 years in industry inside startups, research institutes, antenna manufacturers, and operators, designing a wide variety of products including cellular antennas, digital repeaters, radio units, MRI, NFC, MIMO, and base station RF systems. He designed large-scale antenna systems for 4G/5G as a research director at CMRI. He has more than 1400 citations for over 30 patents and 20 journal papers, and is the Technical Program Co-Chair for IEEE MTT IWS 2015. His research interests are miniature antennas, active antenna arrays, LSAS, metamaterials, massive MIMO, and sensor arrays.

Point-to-Multipoint In-Band mmWave Backhaul for 5G Networks

Rakesh Taori and Arun Sridharan

ABSTRACT

Cost-effective and scalable wireless backhaul solutions are essential for realizing the 5G vision of providing gigabits per second anywhere. Not only is wireless backhaul essential to support *network densification based on small cell deployments*, but also for supporting *very low latency inter-BS communication* to deal with intercell interference. Multiplexing backhaul and access on the same frequency band (*in-band* wireless backhaul) has obvious cost benefits from the hardware and frequency reuse perspective, but poses significant technology challenges. We consider an in-band solution to meet the backhaul and inter-BS coordination challenges that accompany network densification. Here, we present an analysis to persuade the readers of the feasibility of in-band wireless backhaul, discuss realistic deployment and system assumptions, and present a scheduling scheme for inter-BS communications that can be used as a baseline for further improvement. We show that an in-band wireless backhaul for data backhauling and inter-BS coordination is feasible without significantly hurting the cell access capacities.

INTRODUCTION

The next generation of cellular systems (5G) is expected to meet the rapidly growing demands for data through densification of networks using high-capacity small cells [1]. As networks become dense (through *small cell deployments*):

- The cost of bringing fiber to every small cell becomes prohibitively expensive [2].
- Intercell interference increases and becomes a limiting factor in achieving higher cell capacities [3].
- Handover is more frequent, causing the associated handover signaling overhead to increase [4].

Using wireless technologies to lower the cost of *wiring* (digging/trenching to lay wire/fiber in the ground) a cell, for instance, by deploying point-to-point microwave links, is already a well accepted approach [5]. However, even to meet 4G cellular data rate requirements, the number of point-to-point microwave links is already very high [6, 7], and 5G networks are likely to be even more dense. The *first identifiable issue* therefore is *a cost-effective method for connecting 5G BSs to the network* (base station, BS, to network communication).

Intercell interference management necessarily involves inter-BS communication. Extensive studies in Third Generation Partnership Project (3GPP) working groups have revealed that latency is a particularly sensitive issue for the achievable performance of cell edge throughput and/or interference mitigation schemes such as coordinated multipoint (CoMP) [8]. *The second issue* in densification therefore is *low-latency inter-BS communication* (BS to BS communication).

The third issue in network densification is reducing the number of handovers. To reduce the handovers in a dense small cell network, ongoing studies consider user-centric virtual cells such as Cloud Cell [9] *requiring very low-latency communication between BSs* (BS to BS communication).

In summary, providing *low latency and cost effective backhauling mechanisms for connecting 5G BSs* to the other 5G BSs and to the network forms the key challenge for dense deployment of 5G networks.

Here, we consider a point-to-multipoint (PMP) architecture as the basis for a scalable approach to fundamentally address the cost and latency issues described above. In particular, a point-to-multipoint approach will lower:

- *Per-link hardware costs* compared to a dedicated point-to-point link approach. Significant savings in the annual total cost of ownership (TCO) of up to 50 percent can be achieved using the PMP architecture [7].
- *Inter-BS communication latency*. BSs with point-to-multipoint capability can communicate with multiple BSs directly (creating a mesh topology wherein *not all* the links are simultaneously active, but *any* link can be established on demand (cost effective compared to n^2 dedicated links to interconnect n BSs).

To further lower the costs, we consider an *in-band* approach. The term in-band here means that the access link (BS to mobile station, MS, link) as well as the backhaul link (BS-to-BS links or BS to access gateway, AGW, links) are multiplexed on the same frequency band. The advantages of an in-band approach are:

- Cost to the operator of buying separate frequency licenses for backhauling is eliminated (facilitates spectrum reuse).
- A single radio unit is able to serve the backhaul link as well as the access link (facilitates hardware reuse).

Rakesh Taori and Arun Sridharan are with Samsung Research America.

0163-6804/15/$25.00 © 2015 IEEE

The combination of availability of large bandwidths, advanced RF beamforming capabilities using high-gain advanced antennas, as well as MIMO digital beamforming, makes the mmWave cellular system viable for an in-band backhaul solution.

In the current wireless backhaul market [6, 7], out-of-band solutions are prevalent. Even with the higher efficiencies offered by the 4G/Long Term Evolution-Advanced (LTE-A) system, out-of-band solutions are chosen due to the extreme capacity crunch experienced by operators in the expensively licensed frequencies used for access [10, 11]. At millimeter-wave (mmWave) frequencies, however, much wider channel bandwidths can be deployed, and an in-band backhaul approach becomes particularly attractive. The combination of the availability of large bandwidths, advanced RF beamforming capabilities using high-gain advanced antennas, as well as multiple-input multiple-output (MIMO) digital beamforming, makes the mmWave cellular system viable for an in-band backhaul solution. Accordingly, in this article, we consider the use of mmWave frequency bands for cellular access [12] and backhaul.

This article is organized as follows:

- First, we provide a persuasive analysis to establish that an in-band backhaul solution is feasible.
- Next, we discuss the deployment assumptions and system design constraints applicable in a practical system.
- Finally, we present scheduling mechanisms for realizing inter-BS communication and the multiplexing of backhaul links (BLs) and access links (ALs) based on the aforementioned assumptions and deployment scenarios.

Feasibility of In-Band Backhaul

The term in-band, as mentioned in the Introduction, means that the access link (BS-MS link) and backhaul link (BS-BS links or BS-network links) are multiplexed on the same frequency band.

In contrast to the commonly deployed out-of-band wireless backhaul solutions (where dedicated frequency resources can be used for backhaul and access separately), in an in-band wireless backhaul system, the resources used for backhaul are *taken from* the AL resources regardless of the multiplexing method used. An important task, therefore, is to analyze whether enough resources are available for backhauling without compromising the access needs. We begin by setting up the requirements before discussing the feasibility. In the analysis provided below, we assume a typical cellular deployment set up wherein a single BS deployment site consists of multiple (typically three) sectors.

If α is the fraction of resources used for backhauling, an in-band solution is considered feasible if

$$R_b\alpha \geq R_a\,(1-\alpha), \tag{1}$$

where R_a denotes the sum *data rate* of the access links in all sectors and R_b denotes the *sum data* rate of all the backhaul links supporting the cell. This condition can be further expressed in terms of the backhaul and the access link capacities as

$$C_b\alpha\beta_b \geq C_a\,(1-\alpha)\,\beta_a, \ldots \tag{2}$$

where C_a, and C_b denote the maximum uplink (UL) and downlink (DL) access capacity of a BS, respectively, while β_b is the efficiency on the BL, and β_a is the efficiency on the AL.

We analyzed and verified the achievable link capacities through Monte Carlo simulations. Figure 1 summarizes the overall framework used for the simulations. Figure 1a shows the deployment scenario where an inter-site distance (ISD) of 200 m between BSs is considered. Mobile stations (MSs) are dropped randomly at a distance (5, 100) m from the BS within a (–24°, 24°) angular region. The parameters used in the simulation are tabulated in Fig. 1c. Specifically, we consider an mmWave beamformed system at 28 GHz with uniform linear antenna arrays (multiple subarrays) at both the BS and MS sides. We assume 2 × 2 subarrays for the BL (BS–BS link) and 2 × 1 subarrays for the AL (BS–MS link). The BS and MS transmit powers are assumed to be 43 dBm and 23 dBm, respectively. The antenna gains are based on current practical antenna capabilities (23 dBi and 6 dBi at the BS and the MS, respectively). For the BS beams, we assume a uniform angle space between (–30°, 30°) and a sector size of 60°, while a uniform angle space between (–45°, 45°) is assumed for the MS beams. We have considered two scenarios:

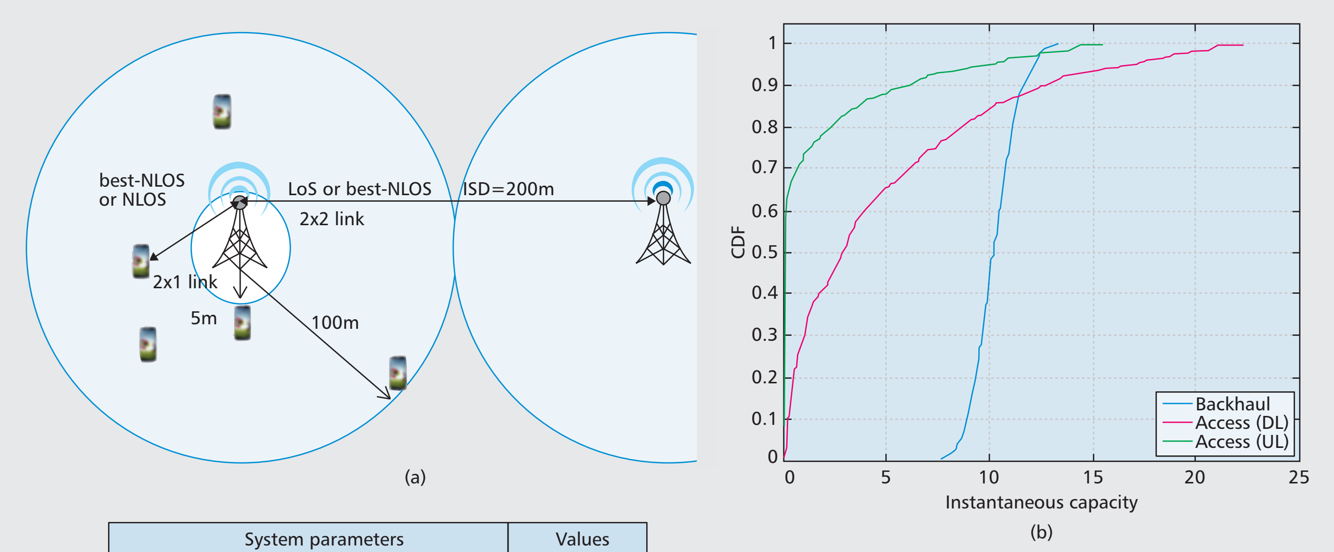
1. Line of sight (LoS) for the BL and best non-LoS (best NLoS) for the AL
2. Best NLoS for the BL and just NLoS for the AL

Best NLoS refers to the best NLoS path among all the NLoS paths. For the results reported here, we have only considered option 2, which is much closer to a real deployment in that mobility constraints will prohibit the best NLoS path from being used for ALs all the time, while a BL can be assumed to be best NLoS even in practical deployments (due to the static nature of the BSs and largely static nature of the reflectors).

Based on the deployment scenario and simulation parameters mentioned above, we obtained the cumulative distributed functions (CDFs) of the instantaneous cell capacities (shown in Fig. 1b). The average link capacities tabulated in Fig. 1d show that the BL capacity is about 2.1 times higher than the AL capacity. Based on these capacity numbers, we can analyze the feasibility. But first let us look at the other factors affecting the BL and the ALs.

Although the resources used for the BL are taken from the AL, we argue that the BL (BS-BS link) is likely to be much more efficient than the AL (BS-MS link) even when the same BS hardware is reused. The main reasons are:

- The BL can be assumed to have high gain antennas on both the Tx and Rx side, while the AL is likely to has fewer antennas on the MS side.
- MS transmit power (uplink) for the AL is much lower; BSs at one or both ends of the BLs can use higher transmit power than the AL.
- Since BLs can be assumed to have fewer channel fluctuations than the AL (which needs to deal with mobility), the BL is likely to incur less overhead than the AL (think of control signaling and reference signals).



System parameters	Values
Channel model	3G SCM UMi
Intersite distance (ISD)	200 m
BS Tx power	43 dBm
MS Tx power	23 dBm
BS antenna gain	23 dBi
MS antenna gain	6 dBi
#Sub-arrays for backhaul link (BS - BS link) #Sub-arrays for access link (BS - MS link)	2 x 2 2 x 1
#Antennas per sub-array at the BS (MS)	8 (2)
#RF beams supported at the BS (MS)	9 (5)

(c)

	Values	Gbits/s
Backhaul	10.25	5.91
Access (DL)	4.8	2.76
Access (UL)	1.59	0.91

(d)

Figure 1. Monte Carlo simulation setup for evaluating wireless backhaul feasibility and observations: a) Monte Carlo simulation scenario for evaluating feasibility of *in-band* wireless backhaul; b) CDFs of the instantaneous capacities on the BL and ALs (uplink, UL, and downlink, DL); c) parameters used for the Monte Carlo simulations; d) average link capacities obtained using Monte Carlo simulations.

- Moreover, BSs are at fixed and higher locations compared to MSs, which also impacts the performance positively.

Since the backhaul link is specifically designed for the wireless backhaul system, we have assumed the PHY overheads in the BL to be about 5 percent (largely static links do not require frequent reference signal transmission, large scheduling information maps [see shceduling scheme description in the next section], etc.). The overhead is assumed to be around 30 percent for the AL (close to the AL overheads observed in cellular systems). Using these values of efficiency, Eq. 2 can be evaluated for different values of α (recall that α is the fraction of resources/bandwidth taken from the AL to serve the BL) and different downlink-to-uplink (DL/UL) ratios. The link capacity results for the best NLoS path on the BL and NLoS path for the AL are captured in Table 1, where "DL/UL" indicates the downlink-to-uplink ratio for resource allocation in a time-division duplex (TDD) system. DL/UL ratios of 1 (symmetric link) up to 3 (DL heavy) are considered on the access link. The first row (link capacity C) in Table 1 shows the link capacity for the BL and AL (DL and UL are shown separately). The second row shows the data rate *R* that is obtainable for the backhaul link (5 percent overhead assumed) and the access link for various DL/UL ratios (assuming ~30 percent overhead). The subsequent rows shows the obtainable data rates for various fractions of α.

It can be observed that using only 25 percent of the AL resources for the BL is sufficient to support user data rates as high as 0.8 Gb/s. Next, we look at the deployment assumptions used in this article.

Deployment Assumptions

We begin with the 4G cellular deployment as a basis. For simplicity, but without compromising the essence, it is assumed in this article that 4G BSs are deployed on a hexagonal grid (as shown in Fig. 2), and that every existing 4G BS is connected to the network using a wired BL (e.g., fiber). It is further assumed that when the much higher capacity 5G deployment commences, the operator is likely to reuse the existing BS sites and utilize the existing wiring *first*, to avoid additional cost of digging/trenching. This will result in the 5G BSs providing much higher cell capacities, but the radio coverage is very likely to shrink (as shown for illustration purposes in Fig. 2).

(Unit: Mb/s)	BL (best NLoS)	AL (NLoS)		
		DL/UL = 1	DL/UL = 2	DL/UL = 3
Link Capacity *C*	3,301	DL 1,817/UL 343		
Data Rate *R*	3,136	784	968	1,061
Data (α = 1/5)	627	627	775	849
Data (α = 1/4)	784	588	726	796
Data (α = 1/3)	1,045	522	646	707

Table 1. Feasible AL and BL rates assuming non-line-of-sight (NLoS) conditons on the AL and best NLoS conditions on the BL for various downlink to uplink ratios. Best NLoS condition can be assumed on the BL as there is ample time to choose the best NLoS setting (which does not change very often) compared to the AL.

To fill these coverage gaps, the operator will need additional 5G BS sites. For backhauling these additional sites, the operator has three choices:

1. Bring wire/fiber to these additional sites (incurs additional costs for digging/trenching).
2. Connect the new 5G BSs to the 5G BSs at the old sites using wireless backhaul ("no additional fiber" scenario).
3. A combination of 1 and 2.

These choices result in different fiber densities (being most dense for option 1 and least dense for option 2), the cost being directly proportional to the density. Assuming the availability of fiber backhaul to the original 4G cell sites and no additional fiber backhaul to the additional 5G sites, depending on how the cells are stacked, the resulting fiber density can either 1/3 (i.e., one of every three 5G cells is fibered), as shown in bottom half of Fig. 3a, or the resulting fiber density can be 1/4 (i.e., one of every four 5G cells is fibered), as shown in the upper half of Fig. 3a. It should be clear that the density can be raised to an arbitrarily higher ratio (as desired) by bringing fiber to some of the unwired 5G cells. In the extreme case, where fiber is brought to all 5G BSs, the resulting density = 1, and this is equal to choice 1 mentioned above.

Here, we focus on the "no additional fiber" scenario (i.e., choice 2), where the coverage of each 5G cells shrinks to about 57.7 percent of the original 4G cell. The additional 5G cells are served solely by wireless backhaul provided by the 5G BSs at the 4G sites, resulting in a situation where one out of every three 5G cells has wired backhaul. Finally, to complete the cellular deployment scenario description, we assume that every BS cell site has three sectors.

In-Band Backhaul Design Considerations

To avoid interference, one can be tempted to think of using space-division multiplexing (SDM) to serve ALs and BLs simultaneously. However, the *AL-BL interference* from the transmission on the BL to the transmission on the AL, and vice versa, is considerable. Depending on the hardware configurations at the transmitter and receivers, significant degradation in signal-to-interference ratio (SIR) can be observed at the receiving MS on the AL and the receiving BS on the other end of the BL as shown in Fig. 4. The SIR curves of Fig. 4 are obtained at the MS for a BS, equipped with an 8 × 8 phased array antenna transmitting simultaneously on the AL and the BL. It is assumed that the BS is located at the bore sight (0° azimuth steering), while the SIR at the MS is observed for different azimuth steering. When the MS is located along the same azimuth direction as the receiving BS, SIRs as low as 0 dB are observed.

Another challenge is the *self-interference* at a single BS site from an antenna panel in one sector to the other antenna panel serving the other sector. When one panel is set to receive, while the other panel is set to transmit, sophisticated signal processing (interference cancellation) is required to cancel the self-interference.

In this article, we work with a minimal (baseline) hardware configuration that does not require space-division multiple access (SDMA) capabilities or self-interference cancellation capabilities. Furthermore, we assume that at a given BS site, all the antenna panels are set to either transmit or receive. To further reduce hardware requirements, it is assumed that one BS sector has one digital chain and can perform one transmission/reception in one sector at any one time. Point-to-multipoint functionality on the BL is only achieved by adaptively/dynamically steering the beams to different neighboring BSs at different times. Lastly, we assume that the BL transmissions are transparent from an MS perspective (i.e., the MS operations remain unchanged, and existing MSs can be used in the network deploying in-band wireless backhaul).

TDM-Based Scheduling Scheme for In-Band Backhauling

Here, we consider the time-division multiplexing (TDM) approach for serving the backhaul and AL under the aforementioned constraints.

TDM schemes have been used for in-band multihop relays [14, 15]. In the TDM-based scheme for in-band backhauling, the access and BL are time-multiplexed by reserving a portion of the access frame for backhaul transmission/reception, where a backhauled or wired 5G BS (W-BS) transmits backhaul data to another 5G BS that is *unwired* (a U-BS). The U-BS is served by the W-BS such that a fraction of the downlink αT_d, and uplink αT_u, access portions is reserved for in-band backhaul transmissions over the BL. On one hand, increasing α allows more data to be backhauled to/from the U-BS cell, which in turn will result in enabling higher access capacities for the U-BS. On the other hand, high values of α directly reduce the access portion of the frame, thereby reducing the capacity of the wired as well as unwired BS cells. This trade-off makes it particularly nontrivial to gauge the overall gain. The numerical capacity analysis-based results summarized in Table 1 provide an initial insight/engineering feel of how the various DL/UL ratios and the fraction of bandwidth allocated for backhauling (α) impacts the AL/BL throughput and the feasibility of the overall approach. Although the approach may look similar to that adopted in relays, a key difference is the absence of hierarchy. For example, in the in-band relaying scheme described in [14] it is the BS that schedules all the relay stations (RSs) that are attached to it. In contrast, the problem here is scheduling among the BSs.

In discussing the TDM-based scheduling schemes for backhauling, let us first understand what needs to be done for the backhauled or wired BSs (W-BSs) or the unwired BSs (U-BSs) to communicate with peer BSs (which may be W-BSs or U-BSs). These are:

- A U-BS needs to transmit and receive backhaul traffic from one or more W-BSs.
- All BSs need to exchange control traffic with the neighboring BSs for interference coordination or handover coordination.

Figure 3b shows BSs with three sectors, as in typical cellular deployments. One of the 5G BSs (shown as W-BS) is connected to the core network with a wired backhaul and serves as the bandwidth injection point. The other two BSs are unwired BSs (U-BSs) and are shown in Fig. 3b as 5G-BSs 1 and 2. The U-BSs are connected to the network via the wireless backhaul provided by the W-BS. Each sector in the W-BS serves one sector of two other 5G U-BSs. The overall configuration used in the analysis is that each W-BS communicates with six neighboring BSs. If each BS cell consists of three sectors, three BLs (one in each sector) can be served such that when sector 1 of W-BS transmits to sector 1 of 5G-BS 1, at that time sectors 2 and 3 of W-BS transmit over BLs to their associated 5G-BS sectors of the neighboring BSs (as shown in Fig. 3b). The beam direction is then steered to target the other three neighboring cells so that sector 1 of W-BS now transmits to sector 1 of 5G-BS 2, with the other two sectors of W-BS following the same procedure. This enables transmission to all six neighboring BSs. Similarly, sectors 1, 2, and 3 of 5G-BS 1 receive over wireless BLs from their associated W-BSs during the Rx portion of the frame. This "transmit to 3, receive from 3" (Tx-3, Rx-3) scheme is used as a basic building block to describe inter-BS scheduling schemes.

To understand how the entire in-band access and backhaul system will be scheduled, let us look at Fig. 5, using which we can outline a BS-to-BS scheduling algorithm capable of scheduling the BL while maintaining the usual BS-to-MS link (access) scheduling. In the description below, we assume the following:

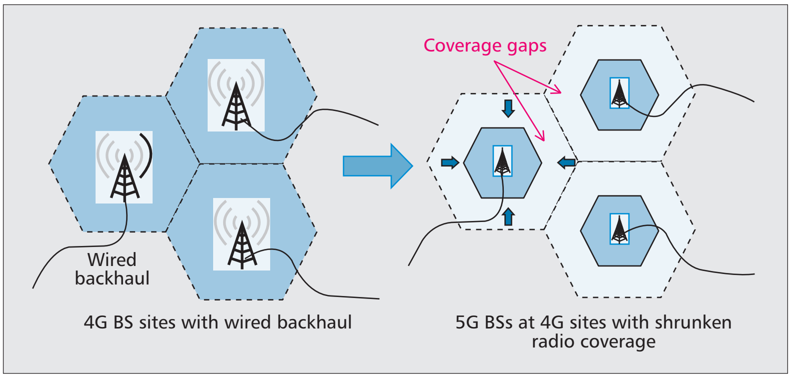

Coverage gaps
Wired backhaul
4G BS sites with wired backhaul
5G BSs at 4G sites with shrunken radio coverage

Figure 2. 5G BSs will be deployed at already backhauled 4G sites.

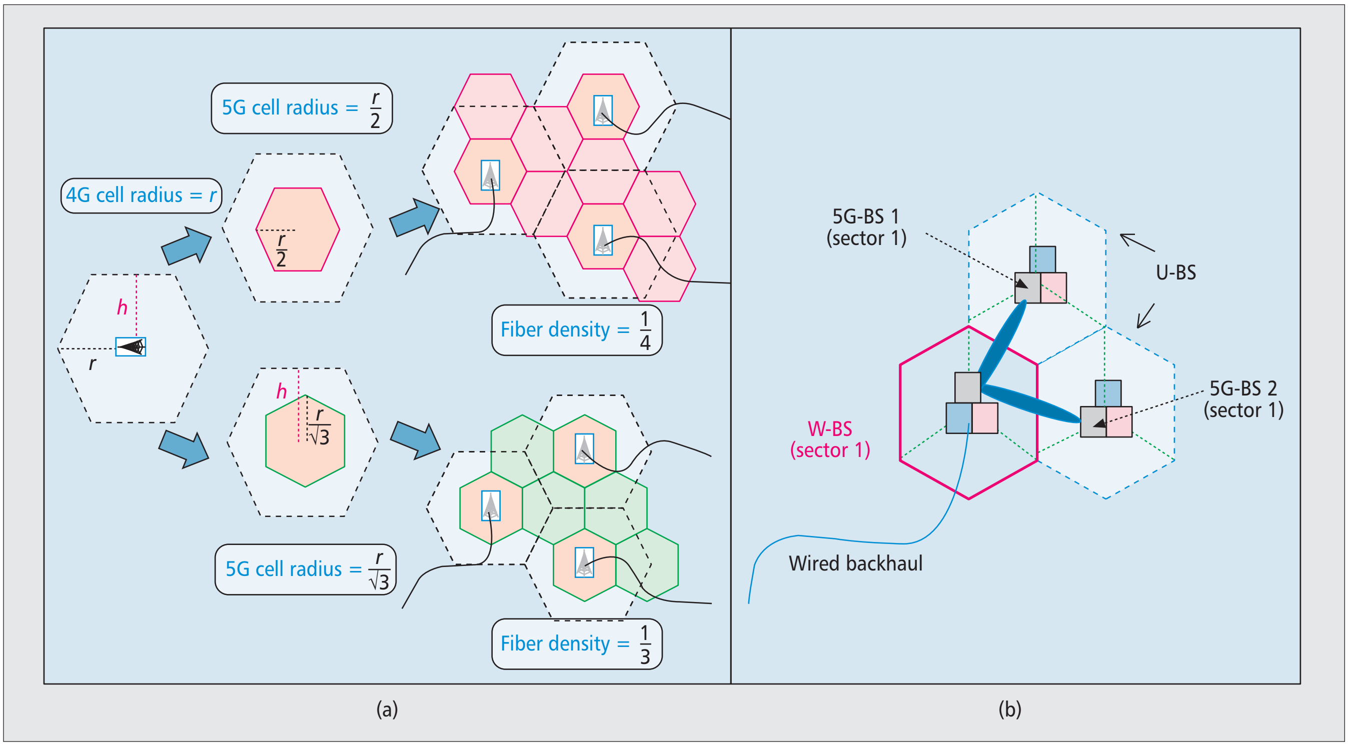


Figure 3. a) Example deployments with different fiber injection densities; b) sector 1 of a W-BS communicates with sector 1 of 5G-BS 1 of an unwired BS (U-BS) in one time slot. Then, in the next time slot, the beam is steered to communicate with 5G-BS 2 (another U-BS).

- A TDD frame structure consists of subframes.
- Each subframe further comprises a DL and a UL portion.
- A few subframes (e.g. 5) constitute a frame.
- Parts of the control signaling are transmitted per frame, while some other control signaling may be transmitted per subframe.

Figure 5 illustrates the mapping of the Tx-3, Rx-3 transmission scheme to a portion of a subframe. The sequence of transmissions can be summarized as:

- Step 1 (shown in the figure): The 5G W-BS transmits to all of its six neighbors, by transmitting to three neighbors at a time using the procedure described above. Since all six 1-tier neighbors of the W-BS are 5G U-BSs, control as well as backhaul data is transmitted to the U-BSs. Note that the transmissions do not require any additional Tx-Rx switching other than the D/L to U/L switch of the TDD systems. Step 1 can, for instance, be scheduled in a portion of subframe 1 of a frame. The remaining portion can be used for data transmission on the AL.
- Step 2: U-BS type 1 (BSs to the right of a W-BS) transmits to all six neighbors. It first transmits only control messages to the three U-BS type 2 neighbors (BSs to the left of W-BS). It then transmits control as well as data to its W-BS neighbors. Step 2 can be accomplished, for instance, in a portion of subframe 2.
- Step 3: U-BS type 2 transmits to all six neighbors. Step 3 can be accomplished, for instance, in a portion of subframe 3.

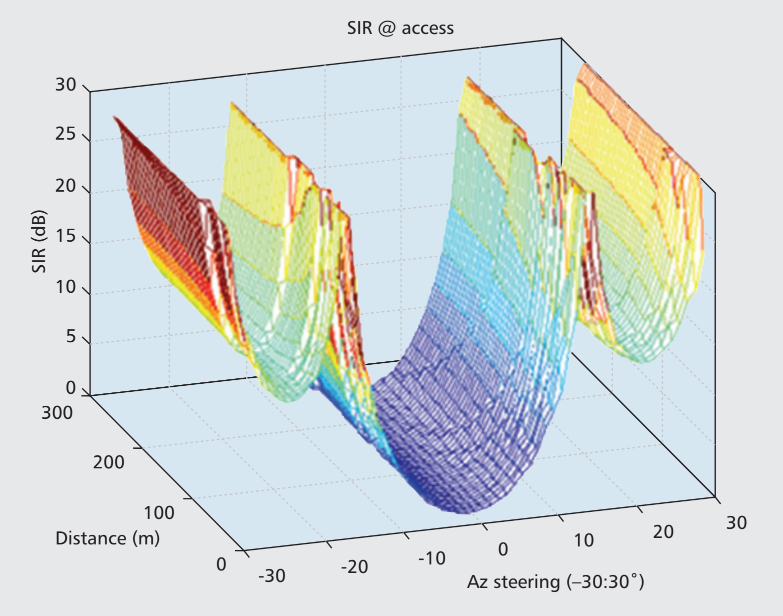


Figure 4. SIR as a function of azimuth beam when the BL is scheduled simultaneously with the AL.

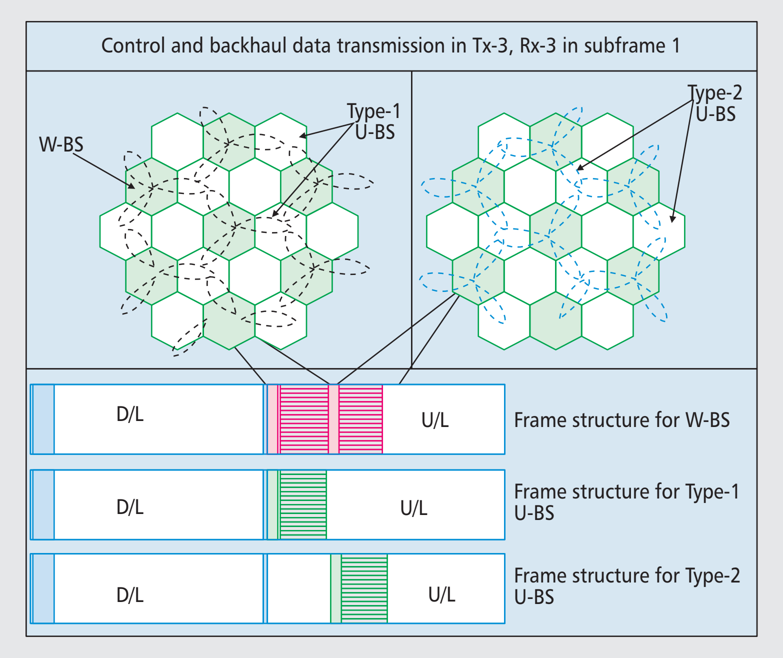

Control and backhaul data transmission in Tx-3, Rx-3 in subframe 1
W-BS
Type-1
U-BS
Type-2
U-BS
D/L
U/L
Frame structure for W-BS
D/L
U/L
Frame structure for Type-1
U-BS
D/L
U/L
Frame structure for Type-2
U-BS

Figure 5. Tx-3, Rx-3 BS-to-BS scheduling scheme mapped to the frame structure for BSs.

If scheduled as described above, by the end of the three subframes, all BSs finish exchanging one round of control messages with their neighbors. Each of the unwired BSs have transmitted and received backhaul data over only a fraction of the subframe. In practice, the backhaul data directed into the unwired BS is much higher, often twice the backhaul traffic leaving the unwired BS to the core network. To increase the backhaul throughput, a fraction of the UL portion of subframes 4 and 5 can be used exclusively for delivering backhaul traffic to the unwired cells.

The scheme outlined above can serve as a baseline scheduling mechanism in that very minimal hardware capabilities are assumed. Performance on the backhaul can be further improved if the transmission and receive equipment at the BS can be assumed to support advanced functions such as SDMA and self-interference cancellation. Improved performance will in turn improve the attractiveness of the in-band solution.

Concluding Remarks

Cost-effective and low-latency solutions for wireless backhaul (including BS-to-network and BS-to-BS communication) will be essential for supporting the envisaged densification in high-capacity 5G networks. In this article, we have presented an initial analysis and solution framework for supporting an in-band, point-to-multipoint, non-line-of-sight, mmWave backhaul.

We have shown using simulations that an in-band solution is feasible at mmWave frequencies for tolerable losses in access capacities and assuming modest hardware capabilities (does not require SDMA, neither full duplexing nor multiple RF chains per sector). BS-to-BS scheduling is not a trivial matter in that unlike BS-to-MS scheduling or BS-RS scheduling [14], there is no established hierarchy among BSs. In this article, we have outlined a BS-to-BS scheduling scheme for the BL that can be multiplexed with the usual BS-to-MS scheduling for the access, and can be considered for an in-band backhaul and access deployment scenario.

Future work should look into detailed system-level simulations for analyzing the performance of the proposed system. Other avenues of interest are investigations into SDMA and full-duplexing capabilities for further spectral efficiency enhancements.

Acknowledgments

The authors are grateful to their colleagues at Samsung Research America for their valuable feedback and discussions. The authors would like to especially thank Md. Saifur Rahman (for Monte Carlo simulations) and Sridhar Rajagopal (for SIR computations).

References

[1] A. Bleicher, "A Surge in Small Cell Sites," *IEEE Spectrum*, Dec. 2012.

[2] "Crucial Economics for Mobile Data Backhaul," An Analysis of the Total Cost of Ownership of Point-to-Point, Point-to-Multipoint, and Fiber Options, by Senza Fili Consulting, 2012

[3] A. Damnjanovic *et al.*, " A Survey on 3GPP Heterogeneous Networks," *IEEE Wireless Commun.*, 2011.

[4] J. G. Andrews, "Seven Ways that HetNets Are a Cellular Paradigm Shift," *IEEE Commun. Mag.*, Mar. 2013.

[5] "Macrocell Mobile Backhaul Equipment and Services Market Share and Forecast Report," *Infonetics Research*, 2012.

[6] "Spectrum-and-Tech-Issues-for-Microwave-Backhaul in Europe," White Paper, Cambridge Broadband Networks, 2010.

[7] "Rethinking Small Cell Backhaul: A Business Case Analysis of Cost-Effective Small Cell Backhaul Network Solutions," white paper by Wireless 20/20.

[8] 3GPP TSG RAN WG1 Meeting #66, R1-112340, "Consideration of X2 backhaul for CoMP.

[9] R. Taori, B.Y. Chang, et. al., "Cloud Cell: Paving the Way for Edgeless Networks in 5G," *Proc. IEEE GLOBECOM*, 2013.

[10] "A Practical Look at LTE Backhaul Capacity Requirements," Maravedis Market Research and Analysis, 2011.

[11] P. J. Pietraski, "The Bandwidth Crunch: Can Wireless Technology Meet the Skyrocketing Demand for Mobile Data," IEEE LISAT, 2011.

[12] Z. Pi and F. Khan, "An Introduction to mmWave Mobile Broadband Systems," *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 101–07.

[13] T. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," *IEEE Access*, May 2013.

[14] J. Sydir and R. Taori, "An Evolved Cellular System Architecture Incorporating Relay Stations," *IEEE Commun. Mag.*, vol. 47, no. 6, June 2009, pp. 115–21.

[15] O. Oyman, J. N. Laneman, and S. Sandhu, "Multihop Relaying for Broadband Wireless Mesh Networks: From Theory to Practice," *IEEE Commun. Mag.*, June 2007.

Biographies

Rakesh Taori (rakesh.taori@samsung.com) is a senior director with Samsung Research America, Dallas, Texas, responsible for research and standardization of next generation communication systems. He is a highly accomplished technology professional with over 22 years of experience in leading cutting edge research and standardization, and a prolific inventor with 75 granted U.S. patents. He has excelled in multiple leadership roles including principal researcher roles with Ericsson Telecommunications and Philips Electronics. He joined Samsung Electronics in 2004 in Suwon, South Korea, responsible for technology development and standardization of 4G (IMT-Advanced) systems. Significant contributions include creation of the industry's first cellular standard for relay (IEEE 802.16j), first multi-hop mesh standard (IEEE 802.11s), and 4G standardization efforts in IEEE 802.16m, which was adopted as one of the IMT-Advanced systems. He was Vice Chair of the IEEE 802.16 Working Group. He has been a leading figure in Samsung's 5G efforts right from inception, driving the development of core MAC and network aspects, introducing core concepts such as cloud cell and in-band backhaul/access, and delivering fundamental technologies for beamformed MAC, which are essential for enabling mmWave band operation. He also serves on the Board of Directors of the Wi-Fi Alliance.

Arun Sridharan (arunsri@gmail.com) received his B.S. degree in electrical engineering from Anna University, Chennai, , in 2005, and M.S and Ph.D. degrees from The Ohio State University (OSU), Columbus, in 2008 and 2012, respectively. He is currently an architect at Akamai Technologies, Cambridge, Massachusetts. His research interests include wireless communication, computer networks, and Internet performance evaluation with emphasis on link scheduling and protocol design.

Future work should look in to detailed system level simulations for analyzing the performance of the proposed system. Other avenues of interest are investigations in to SDMA and full-duplexing capabilities for further spectral efficiency enhancements.

Ultra-Dense Networks in Millimeter-Wave Frequencies

Robert Baldemair, Tim Irnich, Kumar Balachandran, Erik Dahlman, Gunnar Mildh, Yngve Selén, Stefan Parkvall, Michael Meyer, and Afif Osseiran

Robert Baldemair, Erik Dahlman, Gunnar Mildh, Yngve Selén, Stefan Parkvall, and Afif Osseiran are with Ericsson AB, Ericsson Research.

Tim Irnich and Michael Meyer are with Ericsson Germany.

Kumar Balachandran is with Ericsson Inc.

ABSTRACT

Demands for very high system capacity and end-user data rates of the order of 10 Gb/s can be met in localized environments by Ultra-Dense Networks (UDN), characterized as networks with very short inter-site distances capable of ensuring low interference levels during communications. UDNs are expected to operate in the millimeter-wave band, where wide bandwidth signals needed for such high data rates can be designed, and will rely on high-gain beamforming to mitigate path loss and ensure low interference. The dense deployment of infrastructure nodes will make traditional wire-based backhaul provisioning challenging. Wireless self-backhauling over multiple hops is proposed to enhance flexibility in deployment. A description of the architecture and a concept based on separation of mobility, radio resource coordination among multiple nodes, and data plane handling, as well as on integration with wide-area networks, is introduced. A simulation of a multi-node office environment is used to demonstrate the performance of wireless self-backhauling at various loads.

INTRODUCTION: VISION OF THE FUTURE

The world has witnessed the development of four generations of mobile wireless systems: the era of analog cellular telephony led to three further generations of digital systems, symbolized by GSM/EDGE, UMTS/HSPA, and LTE. The 4G/LTE systems in service today have brought wireless Internet access with data rates of tens of Mb/s to hundreds of millions of users. The demand for more data capacity and throughput shows no signs of abating, and can continue to be met if adequate spectrum can be identified for the next chapter toward continuing expansion of mobile technologies.

The next step in this evolution will lead us to a fully Networked Society with a vision of practically unlimited access to information and sharing of data available anywhere and anytime to anyone and anything. This Networked Society goes well beyond ubiquitous access to speech and mobile broadband services to include wireless connectivity for any kind of device that can benefit from being connected, such devices being embedded within vehicles, home appliances, industrial equipment, etc. Technology in the Networked Society will have to support wireless access applications and usage scenarios with a much wider range of characteristics and requirements than the systems of today. As a few examples, future wireless access should allow for:

- General availability of user data rates in the order of 100 Mb/s and above in urban and suburban environments.
- Data rates exceeding 10 Gb/s for ultra-high-speed mobile broadband access, combined with access-network latencies of the order of 1 ms, in specific scenarios.
- Connectivity of a huge amount of "machine" devices optimized for low cost and long battery life time, often with lower requirements on data rates and latency than many other device types.
- Connectivity for mission critical machine devices typically having very high requirements in terms of reliability and low latency.
- Device cost and energy consumption significantly lower than today to enable new types of devices and corresponding applications.
- Usage of spectrum with varying degrees of quality.

5G networks are expected to be deployed with topologies that go beyond the hierarchical base-station-terminal topology of today's cellular networks. This includes:

- Direct connectivity between devices, individually or in clusters, for specific applications such as public safety as well as more general proximal communications.
- Wireless mesh connectivity of clusters of wireless devices in close proximity to each other, with the ability to impose and modify routing topologies on these clusters.
- The ability to deploy wireless backhaul solutions that are capable of aggregating large numbers of remote radio elements with the network.

It is recognized that the wide range of requirements/characteristics outlined above cannot be efficiently met by a single all-encompassing wireless access technology. However, it is of interest that the available options for wireless access are designed from a common framework that can be specialized for particular scenarios, e.g. a wide

0163-6804/15/$25.00 © 2015 IEEE

range of frequency bands and spectrum regulatory frameworks, varying deployment density, topological variations such as star, relay and mesh connectivity, device-to-device communication capabilities, device complexity limitations etc. The "5G" wireless access network can be expected to comprise a set of well-integrated wireless access technologies that will, together with a software controlled transport infrastructure and associated service clouds, jointly enable the Networked Society [1][2]. Figure 1 illustrates how this will include the evolution of the existing wireless access technologies, especially LTE, and new complementary technologies addressing specific applications and usage scenarios for which the requirements cannot be effectively met by a mere evolution of existing technologies.

Mobile broadband will continue to be a key driver for higher overall traffic and higher achievable end-user data rates. The METIS project [8] has, for example, identified an "*Amazingly Fast*" usage scenario with "instantaneous connectivity" and the availability of data rates up to 10 Gb/s [3]. This may correspond to, e.g., public buildings such as arenas, shops, airports, schools, hospitals, and train stations; private property such as campuses, hotels, and multi-dwelling complexes; outdoor environments such as parks, city centers, and densely populated urban areas; and residential environments. Such scenarios will also be characterized by extremely high traffic demands per area.

Ultra-Dense Networks (UDN) are recognized in the METIS project to fulfill the very high demands on system capacity and achievable end-user date rates in such usage scenarios. UDNs are characterized as networks with very short inter-site distances and capable of ensuring low interference levels during communications; envisioned distances between access nodes range from a few meters in indoor deployments up to roughly 50 m in outdoor deployments, i.e. an infra-structure density considerably higher than the most dense networks of today. UDN performance should thus be optimized for low mobility and should provide graceful degradation as the degree of mobility increases, with eventual fall back to overlaid wide-area coverage.

In what follows, we outline the key requirements and characteristics of such a UDN and discuss the key technology components we believe are needed for its realization.

Key Requirements and Characteristics of UDN

Efficient provisioning of data rates in the order of 10 Gb/s implies the need for sufficiently large transmission bandwidth, in the order of at least hundreds of MHz, and preferably around 2 GHz when using the parameters listed in the following section on millimeter-wave band physical layer structure. We foresee frequencies in the lower part of the millimeter-wave band up to 100 GHz to be of interest.

Propagation characteristics in millimeter-wave bands differ significantly from those at the traditional cellular bands in the lower GHz range, and communication links must rely on antenna directivity to meet the link budget. Transmission along a few narrow beams will create line-of-sight or reflective link geometries with short range. Consequently, a very dense deployment of access nodes in the radio network will be needed. Covering indoor access nodes from outdoors will be challenging, and both fiber connectivity and wireless access through access nodes near windows will be employed.

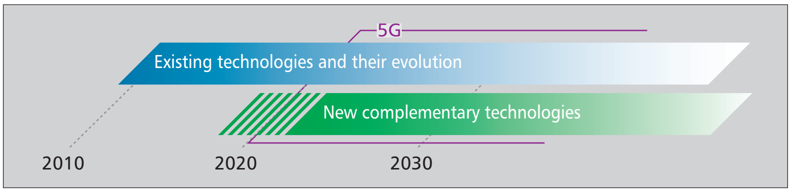

5G
Existing technologies and their evolution
New complementary technologies
2010
2020
2030

Figure 1. 5G is the seamlessly integrated combination of evolved versions of currently existing wireless technologies, such as LTE, and complementary new technologies, jointly enabling the Networked Society for 2020 and beyond.

In very dense deployments, local traffic conditions will fluctuate noticeably. Power levels can also be expected to be relatively balanced between access nodes and terminals. Flexible duplex, in which spectrum resources are dynamically assigned to either transmission direction, is therefore assumed, in contrast to the fixed division of resources currently used by FDD or TDD. Flexible duplex allows up to the full bandwidth to be opportunistically used in each transmission direction. Moreover, flexible duplex can easily exploit unpaired spectrum allocations, which are more likely for large amounts of contiguous spectrum.

Small deployment footprints favor spectrum sharing on an equal basis between unrelated UDN deployments. Wider signal bandwidths at higher frequencies and higher data rates render spectrum efficiency of individual links to be constrained by computational resources and hardware imperfections. However, wider system bandwidths impose a requirement to utilize spectrum efficiently from the point of view of spectrum policy management. Thus sharing between commercial networks can occur, provided that it happens in a controlled fashion. Coordinated management of spectrum and a hierarchical policy enforcement mechanism can ensure a competitive business environment for UDN deployments that need to support high efficiency for spectrum utilization.

UDNs will typically provide local coverage, for example to office buildings, university campus, city centers, etc. It is also important that UDNs are tightly integrated into the overall wireless network architecture with mobility support within the UDN, and between the UDN and a wide-area-covering cellular deployment typically based on LTE.

The high deployment density will need simplified installation procedures and automation of configuration and tuning, as well as connectivity of the UDN to the mobile core. A key enabler is wireless self-backhauling. In other words, full connectivity between UDN access nodes can be provided by the UDN itself.

Some characteristics of a UDN are similar to those of *cellular heterogeneous networks* that incorporate small cells into a macro deployment.

The UDN aims for tight integration with an overlaid cellular network and will support mobility both locally and with the overlaid network. While the UDN can be deployed in the unlicensed 60 GHz band, it will generally support a variety of spectrum bands and regulatory regimes.

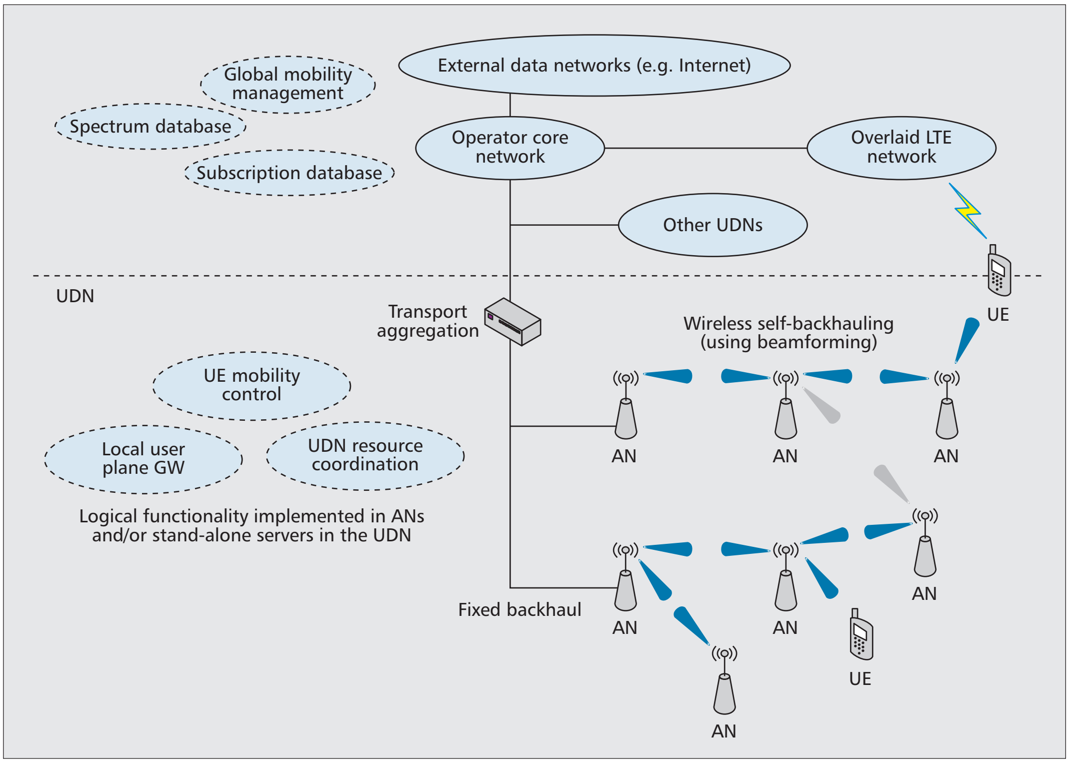

External data networks (e.g. Internet)
Global mobility management
Spectrum database
Subscription database
Operator core network
Overlaid LTE network
Other UDNs
UE
UDN
Transport aggregation
Wireless self-backhauling (using beamforming)
UE mobility control
Local user plane GW
UDN resource coordination
Logical functionality implemented in ANs and/or stand-alone servers in the UDN
AN
AN
AN
AN
Fixed backhaul
AN
AN
UE
AN

Figure 2. Illustration of the UDN architecture.

Indeed, ongoing discussions in 3GPP on certain features, e.g. dual-connectivity, may also be relevant for UDNs. However, a UDN differs from cellular heterogeneous networks in several aspects. The very-wide-bandwidth operation at much higher frequency bands envisioned for UDN imposes system-design limitations, provides opportunities for novel air-interface design, and creates an abundance of resources that may, however, be shared among independent networks. In contrast, operation in frequency bands below 6 GHz as considered for cellular technologies such as LTE focus attention on high spectral efficiency due to the relative scarcity of spectrum.

The intent to deploy UDNs in the millimeter-wave band follows trends in other systems, e.g. IEEE 802.11ad, that specify short range high-bandwidth connectivity in the unlicensed band at 60 GHz. IEEE 802.11ad is mainly aimed toward cable replacement, e.g. wireless HDMI, USB, etc, and supports fast session transfer toward lower-frequency Wi-Fi systems, but is not designed as a mobile networking technology. The UDN aims for tight integration with an overlaid cellular network and will support mobility both locally and with the overlaid network. While the UDN can be deployed in the unlicensed 60 GHz band, it will generally support a variety of spectrum bands and regulatory regimes.

UDN System Concept Outline and Key Features

UDN Architecture and Mobility

The UDN architecture is made up of access nodes (AN) and user equipment (UE). The access node performs scheduling and baseband processing and terminates the radio interface (including physical, medium access, and link layers) toward the UEs. Access nodes can directly be connected to fixed transport or can be wirelessly backhauled toward other access nodes. These physical nodes are aided by several functions responsible for managing various aspects of the UDN including transport and access resource coordination and mobility (see Fig. 2). These functions, described below, are implemented in a virtualized manner that may be distributed within local processing elements such as access nodes or in a centralized facility such as an aggregation router.

UE Mobility Control (UMC): This is specific for each UE and is responsible for controlling intra-UDN UE mobility. The UMC ensures that the UE is anchored to the best access node in its vicinity, collects measurements from UE and neighboring access nodes, manages candidate access nodes capable of communicating with the UE, triggers handover, updates mobility policies in the UE, and manages inactive UE contexts.

Local User Plane Gateway (LUPG): This is UE-specific and handles the user plane. The LUPG anchors local mobility and switches the user plane data path when the UE moves between access nodes. Thus, the LUPG hides local mobility from the operator core network for global services.

UDN Resource Coordination (URC): This UDN-specific functionality is responsible for coordinating resources between access nodes belonging to the UDN, and may implement an inter-UDN resource coordination interface for spectrum sharing. An example of intra-UDN coordination can include interference management, resource assignments for access links, and backhaul links.

Intra-UDN mobility interruption times should reach a maximum of 1 ms with no packet loss. This ensures excellent service experience and high throughput during mobility-induced changes in the communication path. The UMC prepares and activates UDN access nodes that are located in the vicinity of the UE in order to help in evaluating the quality of the neighboring AN-UE path links and to prepare for fast change of links from one access node to another.

Seamless integration of mobility and connectivity with wide-area networks will enable service continuity when the UE loses UDN coverage. Simultaneous connectivity of the UE to the wide-area network can be designed with an efficient sleep mechanism. A fast switching mechanism is used to steer the traffic between the different accesses. This mechanism will also include functionality to support loss-less packet connectivity.

Parameter	Value
Bandwidth: channel/occupied (MHz)	2000/1843
Subcarrier spacing (kHz)	360
Used sub-carriers	5121
Symbol duration (excluding cyclic prefix) (μ s)	2.778
Cyclic prefix (ns)	347.222
Symbol duration (including cyclic prefix) (μ s)	3.125
Subframe time or scheduling interval (μ s)	100
Raw data rate in Mb/s: single stream, 6 bits/symbol (64 QAM)	9830
Raw data rate in Mb/s: dual stream, 6 bits/symbol (64 QAM)	19661

Table 1. Possible OFDM/DFTS-OFDM parameters for a UDN.

Millimeter-Wave Band Physical Layer Structure

A UDN will use frequency assignments between 500 MHz and 2 GHz per network. Resource partitioning between the two transmission directions will be dynamic and flexible, enabling a node to follow fluctuations in traffic load, as expected in small served areas. It is not important for all nodes to handle large bandwidths, and individual devices will have the ability to operate over smaller bandwidths. Spectrum can thus be partitioned across parts of the network with Frequency Division Multiple Access (FDMA).

OFDM, and precoded variants such as DFTS-OFDM, enable both frequency-domain and time-domain partitioning of resources. Such modulations therefore continue to be a good choice for UDNs. While OFDM offers slightly better link performance at the cost of higher Peak to Average Power Ratio (PAPR), DFTS-OFDM has the advantage of lower PAPR but may require a more complicated system design. For example, Frequency Division Multiplex (FDM) together with DFTS-OFDM requires careful consideration to maintain low PAPR. A low PAPR enables simpler hardware, an important factor at millimeter-wave frequencies.

The subcarrier spacing for the frequency-domain signal should be wide enough to provide sufficient robustness toward Doppler and phase noise in the millimeter-wave band. The subcarrier spacing should also be sufficiently narrow to minimize the fractional overhead due to the cyclic prefix, the length of which is determined by the time dispersion in the channel and path delay variations during rapid link handovers. The design parameters listed in Table 1 meet the Doppler and phase noise requirements for the same level of mobility as LTE up to an operating frequency of 45 GHz, and can handle lower and still acceptable Doppler shifts up to an operating frequency of 100 GHz. There is adequate protection in the cyclic prefix for RMS delay spreads of 140 ns with a 15-20 m range to the access node without explicit timing adjustment. Outdoor deployments may need a longer cyclic prefix and a timing adjustment mechanism during initial access. Clock rates are a multiple of the 30.72 MHz rate used by LTE, possibly favorable for integrated designs. Various transmission bandwidths ranging from 100 MHz to 2 GHz can be implemented by varying the total number of subcarriers. As shown, the maximum supported raw physical layer data rate is close to 20 Gb/s . Multi-antenna schemes that realize beamforming functionality, as described later, improve performance by limiting the observable time dispersion, thus limiting the cyclic prefix.

Beamforming and Antenna Solutions in the Millimeter-Wave Band

With a single isotropic antenna element, the received signal spreads omni-directionally away from the transmitting antenna and received power decreases with increasing frequency due to decreasing element aperture. As an example, the received power at 40 GHz is 26 dB lower than at 2 GHz with omni-directional antennas used at both ends of a line-of-sight link. An antenna at either end of the link that maintains aperture (i.e. maintains its effective area) constant, regardless of frequency, can make this loss irrelevant. Such an antenna is, however, no longer omni-directional but directive. Directive transmit and receive antennas can lead to a total net gain in link budget.

Adaptive antenna arrays can vary directivity dynamically using phase (and sometimes amplitude) adjustments of each antenna element. Such phased arrays can use beamforming to dynamically direct signals along desirable radio paths to add constructively. While beamforming targets improved link margin and isolation of interference, spatial multiplexing can further improve peak throughput in high SINR situations. A UDN will primarily rely on beamforming to meet the link budget at high frequencies. Nominal support of two spatial layers per user is expected.

Multi-user MIMO transmissions can be directed to several users, aided by the interference isolation provided by beamforming. The high spatial reuse enabled by beamforming local-

Centralized solutions do not scale well for larger networks but can deliver superior performance. Centralized routing will need dissemination of routes to physical nodes. Routes would be maintained over at least several milliseconds and then updated only on need.

izes individual links, which enables self-backhauling to reuse resources used for access, allowing aggregation of links.

A challenge is the need for accurate beam selection and tracking. Another is the mitigation of outage events by switching rapidly to alternative paths between the transmitter and receiver. One solution would transmit sounding reference signals into different directions at different time instances and let the receiver feedback the best transmit direction. The number of required time slots is proportional to the number of transmit directions. Another possibility would be to map directions to frequencies rather than time slots, thus reducing the time for beam selection, subject to sufficient link margin to meet coverage requirements.

Beamforming can be implemented with a variety of tradeoffs between complexity, flexibility, and power consumption. Simpler approaches realize wideband beamforming in the time domain either in baseband or RF. At the other extreme, digital beamforming at baseband prior to OFDM modulation allows frequency-selective and spatially-selective resource assignments; beamforming weights are applied per resource assignment and antenna element. This solution is flexible but complex due to the large number of digital to analog interfaces and transceiver chains. Other solutions would be between these extremes.

SELF-BACKHAULING

High spatial reuse and abundance of spectrum in the millimeter-wave band are key to the high capacity and throughput of UDNs. Wireless self-backhauling improves reachability and coverage by easing connectivity between access nodes. While some wireless access technologies, e.g. LTE and Wi-Fi 802.11s, support wireless in-band backhauling, they are not designed to operate in the millimeter-wave band with its challenging propagation conditions.

Access nodes in the UDN are connected to aggregation nodes that are in turn connected to a high bandwidth transport network leading to an operator data center before reaching the Internet. The number of aggregation nodes in the network should scale with the size of the network and the volume of traffic. Each hop of the wireless backhaul increases end-to-end delay, and the number of hops is kept small in typical networks. UEs will typically be able to reach an aggregation node within two or three hops. Routing in wired networks is well known and algorithms such as Bellman-Ford and Dijkstra's are studied solutions that can provide optimality [4]. These algorithms require the routing metric associated with each link to be independent from the routing metrics of other links. In wired networks this condition is typically fulfilled. This condition is often unmet in wireless networks in the presence of interference, making validation of optimality challenging. Sub-optimum algorithms with reasonable complexity deliver very good performance [5], and we briefly summarize them as follows.

•Often a route might need to simultaneously fulfill multiple criteria, e.g. low latency and high throughput. It is difficult to provide a joint metric, even though metrics for individual criteria are simple, that fulfills the requirements of well-known routing algorithms. One sub-optimum solution first establishes routes that optimize throughput alone while ignoring latency, and trims those links that fall below a fraction of the maximum throughput. The second phase establishes the route with the best latency for the trimmed network.

•Resource assignments on individual links can affect interference to other links. One solution is to represent each physical node with several virtual nodes, each associated with a different resource assignment on a link. Routing is performed on this virtual network. Once a route is established, each virtual node is mapped back to its corresponding physical node, thus establishing the route. The virtual nodes determine resource assignments along links associated with the physical node.

These routing algorithms can either be centralized or distributed in implementation. In the first case a central node collects all relevant information and computes the route for the complete network (segment), whereas in the distributed case routing decisions are taken locally with the aid of communication between nodes. Centralized solutions do not scale well for larger networks but can deliver superior performance. Centralized routing will need dissemination of routes to physical nodes. Routes would be maintained over at least several milliseconds and then updated only on need. However, traffic variations can occur faster than routes are updated. Resource assignments must therefore operate over a faster time frame within the Medium Access Control (MAC) layer in a manner that is reactive to instantaneous offered load. The MAC layer is supported in this task by the routing with link-specific information that designates:

- Dedicated resources (resources are exclusively assigned by the routing to a link).
- Prohibited resources (resources not allowed on a link due to excessive interference).
- Shared resources (which can be assigned by the MAC layer, and it is the responsibility of the MAC layer to minimize collisions).

SPECTRUM SHARING

Traditionally, mobile wireless access systems have operated in dedicated spectrum identified for IMT in the ITU-R Radio regulations and exclusively licensed for use by a single operator in a given area. Although we expect licensed spectrum to also remain a main spectrum-assignment approach for UDN, we believe that there are clear values in the introduction of spectrum sharing functionality, as follows.

•Exclusive provisioning of system bandwidths in the order of 2 GHz is challenging for multiple parallel network deployments. Deployments in certain premises may often be isolated from other deployments. The capability to operate multiple UDNs on the same frequency channel allows the potential of access to the entire available bandwidth and improves spectrum utilization.

•Spectrum sharing is feasible for UDNs operating at millimeter-wave frequencies since power levels are lower and deployment is more localized.

•Spectrum sharing functionality will make it simpler and faster to find spectrum for UDNs since UDN deployment and evacuation of other

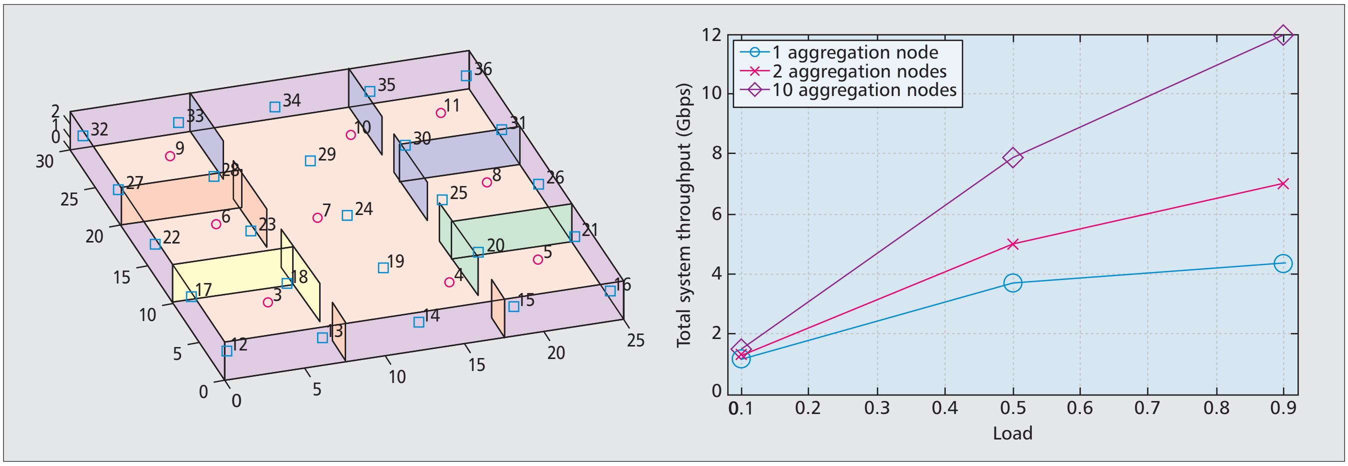


Figure 3. Downlink system throughput of a UDN. Left: Office layout (blue squares = UEs, red circles = access nodes). Distance scales are in meters. Right: System throughput in downlink. Load is the average number of UEs per access node.

services could, if desired, happen simultaneously. Alternatively, services with very different deployment and/or spectrum usage patterns from those of UDNs provide attractive sharing opportunities.

Given the vast range of spectrum sharing approaches discussed in many different constellations, it is of high importance for the 5G community to identify those scenarios that 5G technologies should support. A certain set of scenarios developed in the spectrum sharing toolbox [6] appears relevant for UDNs from a technical and business perspective. From these scenarios, a set of technical enablers can be derived that should become part of 5G technology. At the very minimum, the ability to select an operating channel based on other detected systems by means of dynamic frequency/channel selection (DFS/DCS) is needed.

An inter-UDN coordination protocol will coordinate spectrum sharing based on limited and explicit information exchange with the aim to agree on how to separate interfering transmissions in the time-frequency domain. Another way of handling inter-UDN spectrum sharing in a more centrally controlled manner is through a horizontal spectrum manager that controls spectrum usage among UDNs.

One recently discussed option is to access spectrum under the principle of Licensed Shared Access[1] (LSA) [9]. UDNs are well suited for LSA since they do not require access to spectrum over large areas and in a ubiquitous manner and their overall interference potential is low. The key technical enabler required for LSA is an interface to an external spectrum repository over which a UDN can obtain information on available frequencies and related operational constraints.

The unlicensed band at 60 GHz provides additional bandwidth, around 7 GHz, offering an opportunity for coexistence with unlicensed use, e.g. by using a Wi-Fi coexistence mode [7].

Performance

Simulations were carried out for an office (Fig. 3) with ten access nodes, a varying number of which are aggregation nodes that are connected by fiber, others being wirelessly backhauled. The UDN operates at 60 GHz and utilizes a bandwidth of 1.8 GHz. Transmit power levels of access nodes and UEs are 13 dBm and 10 dBm, respectively. An access node has seven antennas distributed over a hemisphere while each UE has two antennas. A noise figure of 6 dB is assumed. A distributed routing algorithm assigns resources partially; the remaining resources are assigned by the MAC layer. No spatial multiplexing is assumed. For each drop, a user has a full buffer with 20 percent probability and no offered traffic otherwise. The channel is mainly modeled by ray tracing. Random point sources provide statistical variation from a purely specular representation. While self-backhauling has an impact on aggregate throughput, it still provides for a system throughput of around 7 Gb/s, as shown in Fig. 3.

Conclusions

The Networked Society will need evolution of mobile communication beyond the multi-Mb/s mobile broadband services currently enjoyed by billions of people, and will require mobile systems to support a much wider range of requirements than today. The fifth generation mobile networks will be built around an evolution of LTE that will provide wide area coverage, and will be aided by other component access technologies that handle specialized scenarios. Certain localized low-mobility environments will benefit from extremely high data rate access provided by Ultra-Dense Network (UDN) deployments. In this article we define and present the UDN as an important component of 5G wireless access that will boost data rates to the order of 10 Gb/s and lower delays to the order of 1 ms in localized environments. The UDN will operate in the millimeter-wave band, and thereby enable large contiguous bandwidths up to around 2 GHz. The physical layer design can be based on OFDM or DFTS-OFDM, and one design option suitable for a wide range of frequency bands and reasonable mobility is presented. The challenges posed by propagation for receiver performance are met by the use of adaptive beamforming to

[1] LSA refers to the concept of freeing frequency ranges that are underutilized by their incumbent users for additional use.

improve link margin. Self-backhauling together with interference-aware routing is used to wirelessly connect access nodes with a wired aggregation node, thus alleviating the need to physically connect all access nodes to a supporting network. The architecture allows for intra-UDN mobility and enables tight integration with existing cellular standards for mobility, security, and quality of service. Beamforming and short link ranges lower interference, thus making coordinated spectrum sharing possible. Technical solutions have been presented for a representative office environment and performance was validated through simulations.

Beamforming and short link ranges lower interference, thus making coordinated spectrum sharing possible. Technical solutions have been presented for a representative office environment and performance was validated through simulations.

Acknowledgement

The authors would like to thank their colleagues Dennis Hui, Johan Axnäs, and Jonas Kronander for their valuable input and fruitful discussions.

References

[1] Ericsson White Paper on 5G, "5G Radio Access — Research and Vision," June 2013, http://www.ericsson.com/news/130625-5g-radio-access-research-and-vision_244129228_c

[2] R. Baldemair *et al.*, "Evolving Wireless Communications: Addressing the Challenges and Expectations of the Future," *IEEE Vehicular Tech. Mag.*, vol. 8, no. 1, Mar. 2013, pp. 24–30, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6470755.

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

[4] J. Bang-Jensen and G. Gutin, *Digraphs: Theory, Algorithms and Applications*, Springer-Verlag, 2007.

[5] D. Hui and J. Axnäs, "Joint Routing and Resource Allocation for Wireless Self-Backhaul in an Indoor Ultra-Dense Network," PIMRC 2013, London, Sept. 2013.

[6] T. Irnich *et al.*, "Spectrum Sharing Scenarios and Resulting Technical Requirements for 5G Systems," *PIMRC 2013*, London, Sept. 2013.

[7] T. Ihalainen *et al.*, "Flexible Scalable Solutions for Dense Small Cell Networks," WWRF Oulu, 2013.

[8] ICT-317669 METIS project home page https://www.metis2020.com/.

[9] CEPT ECC WG FM 53, "Licensed Shared Access (LSA)," ECC Report 205, Feb. 2014, http://www.erodocdb.dk.

Biographies

ROBERT BALDEMAIR (robert.baldemair@ericsson.com) received his Dipl. Ing. and Dr. degree from the Vienna University of Technology in 1996 and 2001, respectively. In 2000 he joined Ericsson, where he initially was engaged in research and standardization of digital subscriber line technologies ADSL and VDSL. Since 2004 he has been working on research and development of radio access technologies for LTE, and since 2011 with wireless access for 5G. Currently he holds a master researcher position at Ericsson. In 2014 he and collegues at Ericcson were nominated for the European Inventor Award, the most prestigious inventor award in Europe, for their contribution to LTE.

TIM IRNICH holds a diploma and Ph.D. in electrical engineering from RWTH Aachen University. From 2002–2007 he worked as a research engineer at RWTH Aachen University's chair of communication networks and specialized in radio spectrum regulation and management. He joined Ericsson Research in 2007, where he served as project team leader and coordinator of research on technical and regulatory aspects of spectrum sharing and dynamic spectrum access for 5G. He also represented Ericsson in regulatory standardization groups at CEPT and ITU-R. In 2014 he joined Business Unit Cloud & IP at Ericsson, where he is currently working on network function virtualization and cloud topics.

KUMAR BALACHANDRAN has 22 years of experience in the wireless communication industry. He received his bachelor's with honors in 1986 from the Regional Engineering College in Tiruchirapalli, India, and completed his masters and doctorate in computer and systems engineering with a concentration in coding and modulation from Rensselaer Polytechnic Institute, NY, in 1988 and 1992, respectively. He started his career at PCSI, where he helped specify and build Cellular Digital Packet Data (CDPD) and has worked in Ericsson Research for the past 19 years. He is well published, has been invited to speak at several conferences, and served on many panels. He is named on 54 US patents. He holds the position of principal research engineer at Ericsson research, and his current interests are in spectrum and 5G topics.

ERIK DAHLMAN is a senior expert in radio access technologies within Ericsson Research. He was deeply involved in the development and standardization of 3G wireless access. Later on he was involved in the standardization/development of 4G (LTE) wireless access and its continued evolution. He currently focuses on research and development of future 5G wireless access. He is the co-author of the book *3G Evolution — HSPA and LTE for Mobile Broadband* and its follow-up *4G — LTE and LTE-Advanced for Mobile Broadband*. He is a frequent speaker at different international conferences and holds more than 100 patents within the area of mobile communication. In 2009 he received the Swedish Government Major Technical Award for his contributions to the technical and commercial success of HSPA. In the spring of 2014 he and colleagues at Ericsson were nominated for the European Inventor Award, the most prestigious inventor award in Europe, for their contributions to LTE.

GUNNAR MILDH received his M.Sc. in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2000. In the same year he joined Ericsson Research, Ericsson AB, Stockholm, and has since been working on standardization and concept development for GSM/EDGE, HSPA, and LTE. His focus areas are radio network architecture and protocols. He is currently an expert in radio network architecture at the Wireless Access Network Department, Ericsson Research.

YNGVE SELÉN joined Ericsson Research in 2007 after completing his Ph.D. in signal processing at Uppsala University in Sweden the same year. He currently holds a master researcher position at Ericsson and has been involved in future radio access and 5G research for several years, both as an active researcher and project manager.

STEFAN PARKVALL [SM] is currently a principal researcher at Ericsson Research, working on future radio access. He is one of the key persons in the development of HSPA, LTE, and LTE-Advanced, served as an IEEE Distinguished Lecturer 2011 and 2012, and is co-author of the popular books *3G Evolution — HSPA and LTE for Mobile Broadband* and *4G — LTE/LTE-Advanced for Mobile Broadband*. In 2009 he received the Swedish government Major Technical Award for his work on HSPA, and in 2014 he and colleagues at Ericsson were nominated for the European Inventor Award, the most prestigious inventor award in Europe, for their contributions to LTE. Dr Parkvall received the Ph.D. degree in electrical engineering from the Royal Institute of Technology in 1996. His previous positions include assistant professor in communication theory at the Royal Institute of Technology, Stockholm, Sweden, and a visiting researcher at University of California, San Diego, USA.

MICHAEL MEYER received his doctoral degree in electrical engineering from the University of Paderborn, Germany, in 1996, and joined Ericsson Research in Aachen afterward. Until 2008 he held the position of senior specialist for wireless protocol interactions and was actively involved in the concept development and standardization of LTE. Currently he is heading radio network architecture and protocols research at Ericsson Research and is driving 5G research activities.

AFIF OSSEIRAN [SM] is director of radio communications within the Industry Area Telecom at the Ericsson CTO office. He holds a doctorate degree from the Royal Institute of Technology (KTH), Stockholm, Sweden. Since 1999 he has been with Ericsson, Sweden. From November 2012 to April 2014 he managed METIS, the EU 5G flagship project. During 2008-2010 he was the technical manager of the Eureka Celtic project WINNER+. He has published over 50 technical papers in international journals and conferences. He has co-authored two books on IMT-Advanced with Wiley. Afif has co-organized multiple IEEE workshops and served on various technical program committees.

Enabling Device-to-Device Communications in Millimeter-Wave 5G Cellular Networks

Jian Qiao, Xuemin (Sherman) Shen, Jon W. Mark, Qinghua Shen, Yejun He, and Lei Lei

Jian Qiao, Xuemin (Sherman) Shen, Jon W. Mark, and Qinghua Shen are with the University of Waterloo.

Yejun He is with Shenzhen University

Lei Lei is with Beijing Jiaotong University.

ABSTRACT

Millimeter-wave communication is a promising technology for future 5G cellular networks to provide very high data rate (multi-gigabits-per-second) for mobile devices. Enabling D2D communications over directional mmWave networks is of critical importance to efficiently use the large bandwidth to increase network capacity. In this article, the propagation features of mmWave communication and the associated impacts on 5G cellular networks are discussed. We introduce an mmWave+4G system architecture with TDMA-based MAC structure as a candidate for 5G cellular networks. We propose an effective resource sharing scheme by allowing non-interfering D2D links to operate concurrently. We also discuss neighbor discovery for frequent handoffs in 5G cellular networks.

INTRODUCTION

Future fifth generation (5G) cellular networks are being developed to satisfy dramatically increasing data traffic among mobile devices with the emergence of various high-speed multimedia applications [1]. Table 1 summarizes the evolution of cellular networks from 1G to 4G from the aspects of implemented key technologies and the most supported applications. A new generation emerges about every 10 years to significantly improve the transmission rate and support more applications. 5G cellular networks are expected to have much higher network capacity and provide multi-gigabits-per-second data rate for each user to support multimedia applications with stringent quality of service (QoS) requirements. For example, uncompressed video streaming requires a mandatory data rate of 1.78/3.56 Gb/s. These newly emerging bandwidth-intensive applications create unprecedented challenges for wireless service providers to overcome a global bandwidth shortage [2].

Millimeter-wave (mmWave) communication is a very promising solution for future 5G cellular networks. An mmWave communication system has very large bandwidth (multiple gigahertz), which can be translated directly to much higher data rates and overwhelming capacity. Multi-gigabits-per-second transmission at mmWave band has been realized in both indoor (e.g., wireless personal area networks) [3] and outdoor (e.g., wireless mesh networks) systems [4]. The availability of mmWave spectrum and recent advances in RF integrated circuit (RFIC) design motivate industrial interest in leveraging mmWave communication for future 5G cellular networks. MmWave 5G cellular networks are expected to have the main characteristics of highly directional antennas at both wireless devices and base stations, lower link outage probability, extremely high data rate in the widest coverage area, and higher aggregate capacity for many simultaneous users. As a replacement of copper/fiber infrastructure, mmWave mesh networks can be used as a wireless backbone for 5G to provide rapid deployment and mesh-like connectivity.

Generally, device-to-device (D2D) communications provide the connection between two wireless devices either directly or by hopping. D2D communications can be established via the base stations in traditional cellular networks. Specifically, one wireless device needs to communicate with the base station; then the base station conveys the data to another wireless device directly or via backbone networks. Motivated by the increasingly high-rate local services, such as distributing large files among the wireless devices in the same cell, local D2D communications have recently been studied as an underlay to Long Term Evolution-Advanced (LTE-A) 4G cellular networks [5]. It can significantly enhance the network capacity by establishing a path between two wireless devices in the same cell without an infrastructure of a base station. In mmWave 5G cellular networks, local D2D communications can be formed to offload cellular communications, thus supporting more simultaneous users. Meanwhile, global D2D communications can be formed with multihop wireless transmissions via base stations between two wireless devices associated with different cells. Taking advantage of mmWave propagation characteristics and the use of directional antennas, a resource sharing scheme supporting non-interfering concurrent links is

0163-6804/15/$25.00 © 2015 IEEE

Generation	Features	Applications
1G	Deployed in the 1980s. Analog technology.	Voice communication.
2G	Deployed in the 1990s. Digital modulations. Primary technologies are IS-95, CDMA, and GSM.	Voice SMS and low-rate data.
3G	144 kb/s for mobile, 384 kb/s pedestrian, and 2 Mb/s for indoor. CDMA2000, WIMAX, and UMTS-HSPA.	New applications, such as video conference, location-based service.
4G	Require ability of 40 MHz channel with high spectral efficiency. LTE, LTE-A, and IEEE 802.16.m.	Higher rate data, hundreds of megabits per second.

Table 1. Evolution of 1G through 4G cellular networks.

proposed to share network resources among local D2D communications and global D2D communications.

In this article, we focus on building D2D communications over mmWave 5G cellular networks. We discuss the mmWave propagation characteristics and the corresponding challenges to enable D2D communications. The future 5G cellular network architecture and MAC structure are described. A resource sharing scheme to allocate time slots to concurrent D2D links to increase network capacity is proposed. We then conclude the article with a summary and a brief discussion of future work.

MmWave D2D Communications

MmWave Propagation

MmWave communication (with wavelength on the order of millimeters), including the frequency band from 30–300 GHz, has several fundamental propagation features [6]. First, the propagation loss is much higher than that in the microwave band (e.g., 28 dB higher at 60 GHz than at 2.4 GHz) since the free space propagation loss is proportional to the square of the carrier frequency. A high-gain directional antenna is favored to compensate for the tremendous propagation loss and reduce the shadowing effect. Second, the short wavelengths of mmWave bands result in difficulties in diffracting around obstacles. Line-of-sight (LOS) transmissions can easily be blocked by the obstacles. Since non-LOS (NLOS) transmissions in mmWave channels suffer from significant attenuation and a shortage of multipaths, link outage can happen if an LOS link is blocked. Third, mmWave signals have difficulties penetrating through solid materials (e.g., at 40 GHz, 178 dB attenuation for brick wall and over 20 dB attenuation for a painted board). The limited penetration capability could confine outdoor mmWave signals to streets and other outdoor structures, although some signal power might reach inside the buildings through glass windows and wood doors. These propagation characteristics lead to challenges to achieve seamless coverage and reliability [7].

D2D Communications

Enabling D2D communications to handle local traffic can be found in [8], where D2D connections are used for relaying rather than improving the spectrum utilization efficiency. In [9], the traffic loads of the coexisting cellular and ad hoc networks are considered to be independent. Recently, D2D communications used in 4G cellular networks focus on local D2D connections as an underlay to cellular connections. The local D2D communications can reuse the cellular resources to increase spectral efficiency, which has promoted much work in recent years [5].

In mmWave 5G cellular networks, two kinds of D2D communications can be enabled: local D2D communications and global D2D communications. Local D2D communications build the path between two wireless devices associated with the same base station, either directly or by relays if the LOS link between them is blocked. They facilitate the discovery of geographically close devices and reduce the communication cost between these devices. Global D2D communications connect two wireless devices associated with different base stations by hopping via the backbone networks. They include device-to-base-station (D2B) communications and base-station-to-base-station (B2B) communications. In contrast with 4G cellular networks where communications between base stations are performed via fiber links, mmWave communication with a highly directional antenna provides wireless connections with high data rate for B2B communications in mmWave 5G cellular networks.

D2D in mmWave 5G

As described earlier, D2D communications are expected to be an essential feature of mmWave 5G cellular networks, to improve network capacity and build connections between two wireless devices. Due to the directional antenna and high propagation loss, mmWave communication has relatively low multi-user interference (MUI), which can support simultaneous communications. By allowing multiple concurrent D2D links, the network capacity can be further improved.

In mmWave 5G cellular networks, D2D communications may face two kinds of potential interference within each cell: interference among different local D2D communications (if there are multiple local D2D communications) and interference between local D2D communications and D2B/B2B communications. Most of the existing works on D2D communications focus on the design of optimized resource sharing algorithms by managing the interferences [5, 10]. In [5], the performance of frequency reuse among D2D links is analyzed with dynamic data arrival settings to obtain average queue length, mean throughput, average packet delay, and packet dropping probability. In [10], the system aims to optimize the throughput over the shared resources while fulfilling prioritized cellular service constraints. The performance of the D2D underlay system is evaluated in both a single-cell scenario and the Manhattan grid environment. It considers resource sharing between one cellular connection and one local D2D connection.

To the best of our knowledge, previous works on resource sharing for D2D communications consider the mutual interference of omnidirectional antennas. Taking advantage of high propagation loss and the use of directional antennas, more D2D links can be supported in each cell in mmWave 5G networks to further enhance network capacity and improve spectrum efficiency. A new resource sharing scheme considering directional interference is necessary in mmWave 5G cellular networks to enable multiple D2D communications.

Network Architecture

It is expected that the current 4G cellular networks can provide seamless coverage and reliable communications because of the lower frequency band. For smooth and cost-efficient transition from 4G to 5G, 5G cellular networks use the hybrid 4G+mmWave system structure shown in Fig. 1 to achieve seamless coverage and high rate in most coverage areas. The management information and low-rate applications (e.g., voice, text, and web browser) are transmitted in 4G networks, while the mmWave bands are available for high-rate multimedia applications.

The 5G cellular networks consist of 4G base stations, mmWave base stations, and mobile devices. In 4G networks, the whole geographical area is partitioned into cells, each of which is covered by one or more 4G base stations. MmWave transmission/reception is based on high directional antennas, which can greatly reduce the mutual interference between mmWave base stations. It has been proved and demonstrated [4] that for an outdoor environment, the interference among mmWave concurrent links are negligible, and directional mmWave communication links can be considered as *pseudo-wired*. Therefore, mmWave base stations do not need to be deployed in cells. In this article, dense mesh networks are adopted for the mmWave backbone with grid topology deployment to provide high rates and aggregate capacity. As shown in Fig. 2, each wireless device has the communication modes of both 4G operation and mmWave operation, and supports fast mode transition between them. Two devices can communicate with each other in the same mode. This article focuses on enabling D2D communications at mmWave band for 5G networks. Therefore, in the following parts of the article, without special indications, the base station refers to the mmWave base station. All wireless devices and mmWave base stations are equipped with electronically steerable directional antennas for mmWave communication. All wireless devices and 4G base stations have omnidirectional antennas for 4G communications. It is assumed that with mmWave beamforming technologies [11], each transmission pair can determine the best transmission/reception beam patterns for data transmission.

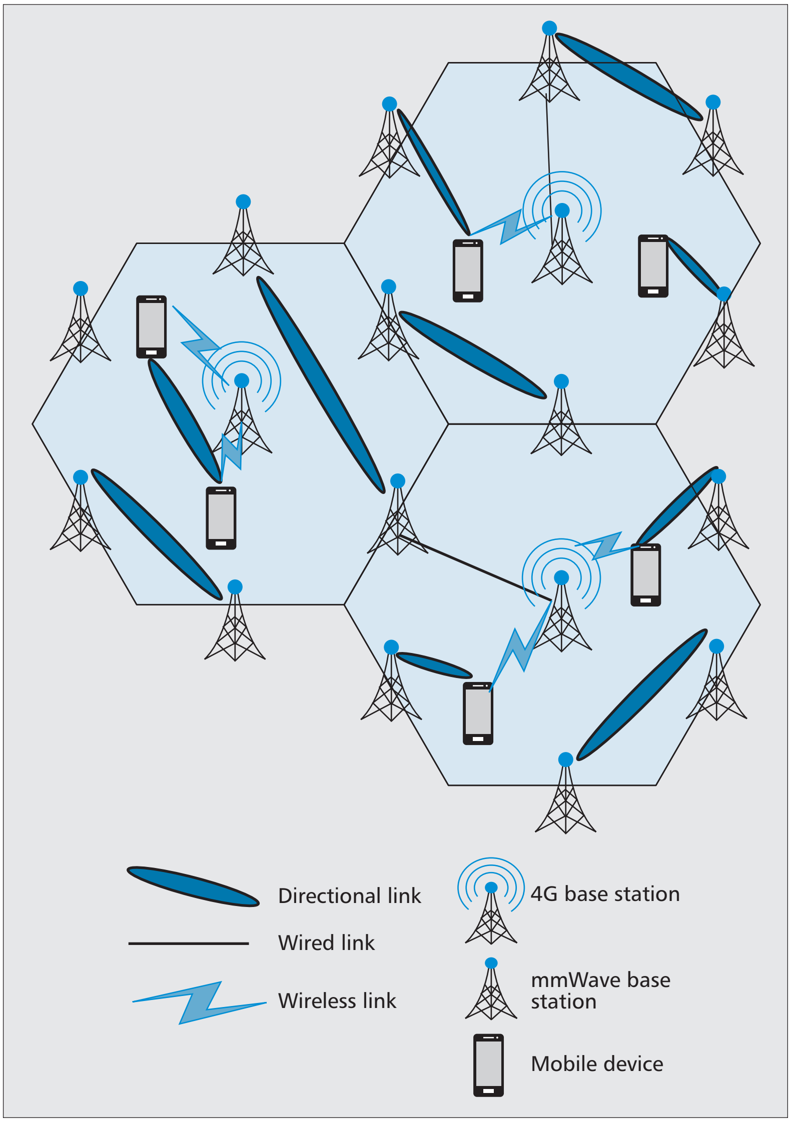


Figure 1. MmWave 5G cellular network architecture.

Medium Access Control

Several works on directional mmWave MAC for networks with low user mobility (e.g., WLAN or WPAN) have appeared in the literature [12, 13]. Cross-layer modeling and design approaches are presented in [12] to account for the problems of directionality and blockage. In the proposed MAC protocol, an intermediate node is randomly selected as the relay if the LOS link between the source and the destination is not available. In [13], an exclusive region (ER)-based resource management scheme is proposed to exploit the spatial reuse, and the optimal ER sizes are derived. The main challenge in mmWave MAC design is how to use the spectrum efficiently to achieve higher capacity considering mmWave propagation features while providing reliable high-rate connections.

MmWave 5G cellular networks support multimedia applications with stringent QoS requirements. To provide guaranteed performance, time-division multiple access (TDMA) is adopted for mmWave channel access in 5G networks with the superframe shown in Fig. 3. Each base station handles the local D2D transmissions, B2B transmissions, and D2B transmissions. Time is partitioned into superframes, each of which are composed of *M* time slots called channel time allocation (CTA). In each CTA, multiple local D2D communications can operate simultaneously to exploit spatial reuse and improve spectrum utilization efficiency. Due to the half-duplex constraint, there should be at most one D2B/B2B link in each CTA since the base station cannot transmit and receive simultaneously. The 4G base stations collect the transmission requests and signaling information for mmWave communication by reliable 4G networks.

For each local communication (including local D2D and D2B), the transmitter polls the receiver to check connectivity. Each receiver has to respond within a fixed interval, that is, a poll inter frame space (PIFS), with a poll response message if the connection is not blocked. The absence of a poll response at the receiver indicates the link blockage and triggers multihop transmission to bypass the obstacles by intelligently selecting a relay within the wireless devices under the control of the base station. Relay selection has great impact on its flow throughput and interference to other links operating at the same time. There are many existing schemes to determine relay selection [3]. Since the main focus of this article is to enable D2D communications, we simplify the relay selection by randomly picking up a node that is close to the direct path of the source and destination with LOS transmissions available to both. The link budget is used to ensure the link reliability over the coverage range. After the transmitter receives the polling response message, it starts to send packets to the receiver. Then the receiver acknowledges the successful packet reception with an ACK message. For transmissions among mmWave base stations, it is assumed that the path can be determined by routing protocol without the involvement of a blocked link in the path.

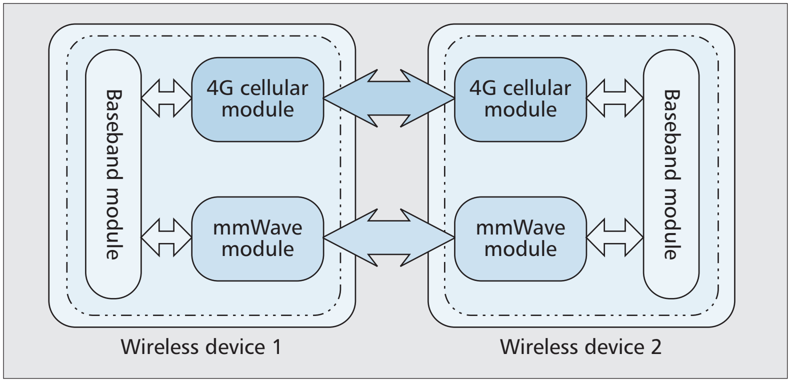


Figure 2. Wireless operation mode of each node.

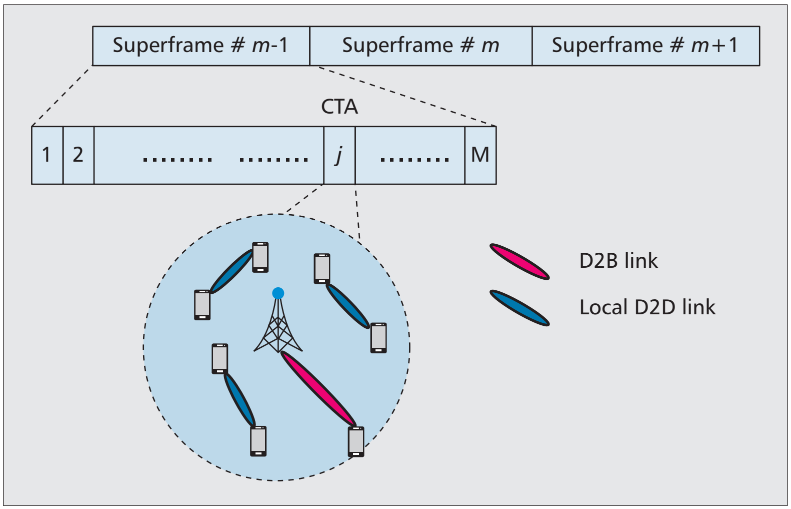


Figure 3. MmWave communication superframe in 5G cellular networks.

Resource Sharing

From the above discussions, resource sharing is the essential problem in enabling concurrent D2D communications in mmWave 5G cellular networks. This section presents the resource sharing modes, formulates the general resource sharing problem in directional mmWave 5G networks, and proposes an efficient algorithm to obtain the resource sharing solution.

Resource Sharing Modes

The local D2D and D2B/B2B links share the resources in mmWave 5G cellular networks. The resource sharing decisions are made by the base station. Generally, there are two resource sharing modes in the network:

- Non-orthogonal sharing (NOS) mode: Local D2D links and D2B/B2B links reuse the same resource, causing interference with each other. The base station coordinates the usage of resources (e.g., transmission power and time slot) for both kinds of links.
- Orthogonal sharing (OS) mode: Local D2D links use part of the resources while the other resources are allocated to D2B/B2B links. Thus, there is no interference between them, which simplifies the resource sharing.

Although orthogonal sharing mode can make resource sharing simple, non-orthogonal sharing can result in better resource utilization efficiency with proper sharing schemes. In this article, the non-orthogonal sharing mode is adopted for multiple concurrent links under the control of the base station. The use of directional antenna and high propagation loss can result in relatively lower mutual interference or even no interference by properly selecting the concurrent links formed by geographically distributed wireless devices.

Some of the existing work on resource sharing of D2D communications consider the scenario of one local D2D and one D2B link to simplify the interference [10]. Concurrent transmissions are also enabled in WLAN/WPAN networks to exploit spatial reuse [14, 15]. These papers consider D2D connections as local communications within the network operated by a network controller. The resource sharing scheme can be either distributively determined by the wireless devices themselves or centrally operated by the base station. As the mmWave 5G cellular networks are centralized in nature, the resource sharing scheme in this article is determined by the base station considering mutual interference among D2B and local D2D connections.

Optimization of Resource Sharing

Due to the long transmission distance and highly directional antennas, the interferences of the concurrent transmissions among mmWave base stations are negligible. The network capacity is mainly constrained by the interferences generated by local network. Each time slot can be allocated to multiple communication links which are spatially separated or overlapped without much interference. Both D2D and D2B/B2B links use the same time slots, and they might interfere with each other. Different sets of active local D2D links may affect the transmission rate of D2B/B2B links and vice versa. How to share the resources among D2D and D2B/B2B links to achieve optimal system throughput is an important and challenging issue.

The resource sharing determines a set of active links for each time slot in the superframe. Total data transmitted in the whole superframe is used as the objective function to achieve the best resource sharing while satisfying the transmission requests of each link. A variable $X_{i,j} = 0$ or 1 indicates if link i is active in the jth time slot. Total data transmitted in the whole superframe can be expressed as the function of $|X_{i,j}|_{L\times M}$ with each rate estimated by Shannon capacity formula. M denotes the number of time slots in each superframe, and L is the number of collected transmission requests.

The above optimization problem is a nonlinear integer programming problem. One possible approach is to relax the integer variables into continuous ones, and use optimization tools to solve the approximated problem. However, the approximated problem is still difficult to solve, since its objective function is not necessarily concave. The complexity of the above problem increases exponentially with the number of concurrent links and number of time slots. In this article, a heuristic resource sharing scheme is proposed to assign a set of active links for each time slot effectively.

Resource Sharing Scheme Design

The complexity of achieving the best resource sharing comes from the possible mutual interference of directional antennas. To simplify the problem and obtain an efficient resource sharing scheme, only non-interfering links are allocated to each time slot to share the resources. The concurrent transmission condition is that two links can operate simultaneously without interference if and only if any transmitter is outside the beamwidth of the other receiver or does not direct its beam to the other receiver if it is within the beamwidth of the other receiver. We apply an ideal "flat-top" model for directional antennas, that is, unit gain within the beamwidth and zero gain outside the beamwidth.

The details of the proposed resource sharing scheme are as follows. By a polling process, if an LOS link is blocked, a relay is selected to build a multihop path. At the beginning of each superframe, all the transmission requests are collected by 4G networks. Transmission requests would be forwarded to mmWave base stations if they require high data rate. The mmWave base station makes the resource sharing decisions for each superframe (i.e., a specific set of active links for each time slot) and sends the decisions to all the involved wireless devices via reliable 4G networks. It is assumed that all the wireless devices and base stations are synchronized.

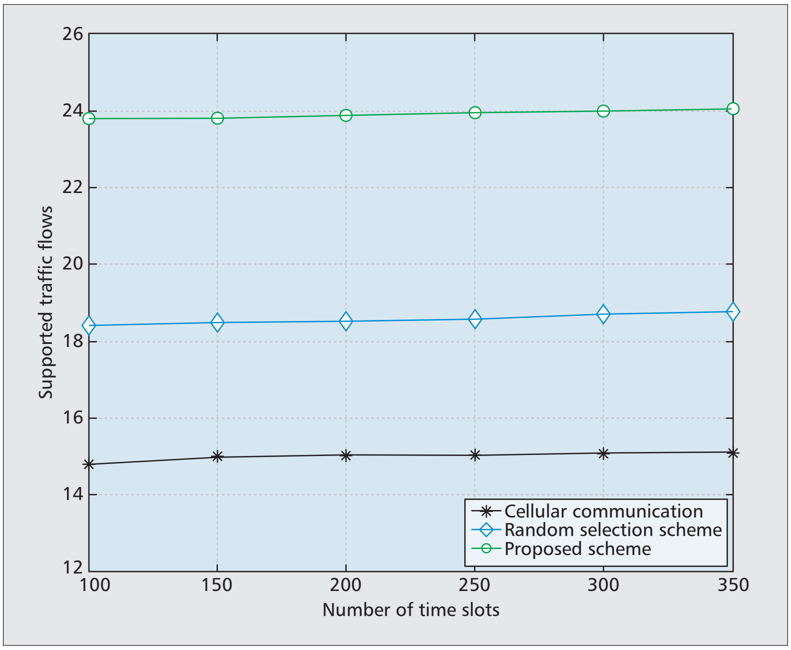


Figure 4. Number of supported traffic flows.

Since the concurrent links rely on LOS transmission, and we allow non-interfering links to operate concurrently, the wireless channel can be modeled by the free space Friis transmission equation. The instantaneous transmission rate can be estimated by the Shannon capacity formula. Each transmission request indicates a minimum average throughput to support multimedia applications. Thus, the number of time slots in each superframe for each transmission request can be predetermined. We randomly sort the transmission links in a specific sequence. A transmission request r_i from the ith link needs $n(i)$ slots. The base station sequentially checks if the ith link can operate concurrently with all the existing links in the same time slot according to the concurrent transmission condition. Note that two links having the same node cannot operate simultaneously due to the half-duplex constraint of wireless communications. If a link does not interfere with all existing links, this link is set to be active in the current time slot. After traversing all the links, the active link set for the current time slot is obtained. This active link set is used for the following time slots until at least one link's throughput requirement is satisfied. If a link's required number of time slots has been satisfied, it should be set inactive, and it is not necessary to check the concurrent transmission condition of this link in the following time slots. The above procedure is repeated until all the time slots have been traversed. If a link's request is not satisfied in the current superframe, it will be re-sent in the next superframe to share the resources with other links.

Figures 4 and 5 show the performance of the proposed resource sharing scheme. There are 40 transmission requests received in the base station. All the wireless devices are randomly distributed in a 20 m × 20 m square area. The transmission

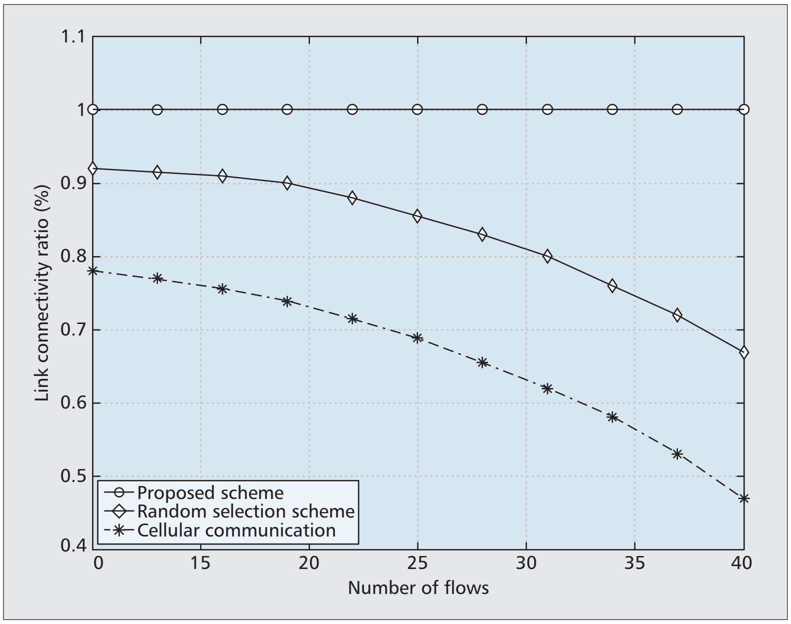


Figure 5. Network connectivity ratio.

power is 0.1 mW, and the background noise level is 134 dBm/MHz. The antenna beamwidth is 45° for both mobile devices and mmWave base stations. The performance of the proposed resource sharing scheme is compared to two other schemes, traditional cellular and random selection. The traditional cellular scheme does not have local D2D communications, while the random selection scheme just randomly selects several links to share the resource. In Fig. 4, the proposed resource sharing scheme significantly outperforms the other two schemes in terms of the number of supported flows by effectively exploiting the spatial reuse opportunities. The proposed resource sharing scheme is very useful, especially for a dense network in the urban area. Network capacity for mmWave 5G networks is an essential issue in the deployment of mmWave base stations. This article considers concurrent transmissions to improve local network capacity.

The proposed resource sharing scheme uses multihop transmission with relays to deal with link blockage. The blockage model defined in the IEEE 802.11ad channel model document is adopted. In mmWave 5G cellular networks, both the obstacles and the mobility of mobile devices can cause link outage if LOS transmission is blocked. Network connectivity is shown in Fig. 5 with various numbers of transmission requests in the network. A relaying mechanism can reduce the link outage probability by replacing a blocked link with an alternative path with two links. The relaying mechanism to keep network connectivity is effective for users with low mobility.

Conclusion and Future Research

In this article, we have discussed the suitability of mmWave band for 5G cellular networks. We have also proposed a resource sharing scheme for concurrent D2D communications in mmWave 5G cellular networks that can significantly improve network capacity while keeping network connectivity well. The article should be useful for future research on enabling D2D communications in mmWave 5G cellular networks.

To achieve high transmission rate and aggregate capacity, mmWave base stations may be densely deployed, especially for urban areas. Thus, mobile users may have to hand off frequently between mmWave base stations. Fast neighbor discovery is required in the handoff procedure for mobile users to find nearby base stations and switch to the base station with better link quality. Although directional antennas offer many advantages on improving spatial reuse and network capacity, there are challenges (e.g., deafness problem) in neighbor discovery. In our future work, we will study neighbor discovery for frequent handoffs with directional antennas in mmWave 5G cellular networks.

References

[1] T. S. Rappaport *et al.*, “Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!,” *IEEE Access*, vol. 1, May 2013, pp. 335–449.

[2] Z. Pi and F. Khan, “An Introduction to Millimeter-Wave Mobile Broadband Systems,” *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 101–07.

[3] J. Qiao *et al.*, “Enabling Multi-Hop Concurrent Transmissions in 60 GHz Wireless Personal Area Networks,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, Nov. 2011, pp. 3824–33.

[4] R. Mudumbai, S. Singh, and U. Madhow, “Medium Access Control for 60 GHz Outdoor Mesh Networks with Highly Directional Links,” *Proc. IEEE INFOCOM*, Apr. 2009, pp. 2871–75.

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

[6] S. Y. Geng *et al.*, “Millimeter-Wave Propagation Channel Characterization for Short-Range Wireless Communications,” *IEEE Trans. Vehic. Tech.*, vol. 58, no. 1, Jan. 2009, pp. 3–13.

[7] K. Zheng *et al.*, “Stochastic Performance Analysis of a Wireless Finite-State Markov Channel,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, Feb. 2013, pp. 782–93.

[8] Y. D. Lin and Y. C. Hsu, “Multihop cellular: A New Architecture for Wireless Communications,” *Proc. IEEE INFOCOM*, Mar. 2000, pp. 1273–82.

[9] K. Huang, V. Lau, and Y. Chen, “Spectrum Sharing Between Cellular and Mobile Ad Hoc Networks: Transmission-capacity Trade-off,” *IEEE JSAC*, vol. 27, no. 7, June 2009, pp. 1–10.

[10] C. H. Yu *et al.*, “Resource Sharing Optimization for Device-to-Device Communication Underlaying Cellular Networks,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, Aug. 2011, pp. 2752–63.

[11] H. H. Lee and Y. C. Ko, “Low Complexity Codebook-Based Beamforming for MIMO-OFDM Systems in Millimeter-Wave WPAN,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, Nov. 2011, pp. 3607–12.

[12] S. Singh *et al.*, “Millimeter Wave WPAN: Cross-Layer Modeling and Multihop Architecture,” *Proc. IEEE INFOCOM*, May 2007, pp. 2336–40.

[13] L. X. Cai *et al.*, “REX: A Randomized EXclusive Region based Scheduling Scheme for mmWave WPANs with Directional Antenna,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, Jan. 2010, pp. 113–21.

[14] J. Qiao *et al.*, “STDMA-Based Scheduling Algorithm for Concurrent Transmissions in Directional Millimeter Wave Networks,” *Proc. IEEE ICC*, June 2012, pp. 1–5.

[16] C. Sum *et al.*, “Virtual Time-Slot Allocation Scheme for Throughput Enhancement in a Millimeter-Wave Multi-Gbps WPAN System,” *IEEE JSAC*, vol. 27, no. 8, Oct. 2009, pp. 1379–89.

Biographies

Jian Qiao (qiaojian1@gmail.com) received his B.E. degree from Beijing University of Posts and Telecommunications, China, in 2006 and his M.Sc. degree in electrical and computer engineering from the University of Waterloo, Canada, in 2010. He is currently working toward his Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include next generation cellular networks, millimeter-wave communication, medium access control, and resource management.

XUEMIN (SHERMAN) SHEN [M'97, SM'02, F'09] received his B.Sc. (1982) degree from Dalian Maritime University, China, and his M.Sc. (1987) and Ph.D. (1990) degrees from Rutgers University, New Jersey, all in electrical engineering. He is a professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo. He was the Associate Chair for Graduate Studies from 2004 to 2008. His research focuses on resource management in interconnected wireless/wired networks, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He served as the Technical Program Committee Chair/Co-Chair of IEEE INFOCOM '14 and IEEE VTC '10 Fall, Symposia Chair of IEEE ICC '10, Tutorial Chair of IEEE VTC '11 Spring and IEEE ICC '08, Technical Program Committee Chair of IEEE GLOBECOM '07, General Co-Chair of Chinacom '07 and QShine '06, Chair of the IEEE Communications Society Technical Committees on Wireless Communications, and P2P Communications and Networking. He also serves or has served as Editor-in-Chief of *IEEE Network*, *Peer-to-Peer Networking and Application*, and *IET Communications*; a Founding Area Editor of *IEEE Transactions on Wireless Communications*; an Associate Editor of *IEEE Transactions on Vehicular Technology*, *Computer Networks*, *ACM/Wireless Networks*, among others; and a Guest Editor of *IEEE JSAC*, *IEEE Wireless Communications*, *IEEE Communications Magazine*, *ACM Mobile Networks and Applications*, and more. He received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, and 2010 from the University of Waterloo, the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, and a Distinguished Lecturer of the IEEE Vehicular Technology and Communications Societies.

Jon W. Mark [M'62, SM'80, F'88, LF'03] received his Ph.D. degree in electrical engineering from McMaster University in 1970. In September 1970 he joined the Department of Electrical and Computer Engineering, University of Waterloo, where he is currently a Distinguished Professor Emeritus. He served as the Department Chairman during the period July 1984–June 1990. In 1996 he established the Center for Wireless Communications (CWC) at the University of Waterloo and is currently serving as its founding Director. He was on sabbatical leave at the following places: IBM Thomas J. Watson Research Center, Yorktown Heights, New York, as a visiting research scientist (1976–1977); AT&T Bell Laboratories, Murray Hill, New Jersey, as a resident cConsultant (1982–1983): Laboratoire MASI, Université Pierre et Marie Curie, Paris, France, as an invited professor (1990–1991); and the Department of Electrical Engineering, National University of Singapore, as a visiting professor (1994–1995). He has previously worked in the areas of adaptive equalization, image and video coding, spread spectrum communications, computer communication networks, ATM switch design, and traffic management. His current research interests are in broadband wireless communications, resource and mobility management, and cross-domain interworking. He is a Fellow of the Canadian Academy of Engineering. He is the recipient of the 2000 Canadian Award for Telecommunications Research and the 2000 Award of Merit of the Education Foundation of the Federation of Chinese Canadian Professionals. He was an Editor of *IEEE Transactions on Communications* (1983–1990), a member of the Inter-Society Steering Committee of IEEE/ACM Transactions on Networking (1992–2003), a member of the IEEE Communications Society Awards Committee (1995–1998), an Editor of *Wireless Networks* (1993–2004), and an Associate Editor of *Telecommunication Systems* (1994–2004).

QINGHUA SHEN received his B.Sc. and Master's degrees in electrical engineering from Harbin Institute of Technology, China, in 2008 and 2010, respectively. He is currently working toward a Ph.D. degree in the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include resource allocation for e-healthcare systems, cloud computing, and smart grid.

YEJUN HE received a Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology in 2005. He is a professor at Shenzhen University. He has been a visiting professor at the University of Waterloo and Georgia Institute of Technology. His research interests include channel coding and modulation, MIMO-OFDM wireless communication, space-time processing, and smart antennas.

LEI LEI received a B.S. degree in 2001 and a Ph.D. degree in 2006, respectively, from Beijing University of Posts and Telecommunications, both in telecommunications engineering. From July 2006 to March 2008, she was a postdoctoral fellow at Computer Science Department, Tsinghua University, Beijing, China. She worked for the Wireless Communications Department, China Mobile Research Institute from April 2008 to August 2011. She has been an associate professor with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, since September 2011. Her current research interests include performance evaluation, quality of service, and radio resource management in wireless communication networks.

Hybrid Millimeter-Wave Systems: A Novel Paradigm for HetNets

Hani Mehrpouyan, Michail Matthaiou, Rui Wang, George K. Karagiannidis, and Yingbo Hua

Hani Mehrpouyan is with California State University.

Michail Matthaiou is with Queen's University Belfast and Chalmers University of Technology.

Rui Wang is with Tingji University.

George K. Karagiannidis is with Aristotle University of Thessaloniki, Greece, and Khalifa University, Abu Dhabi, UAE.

Yingbo Hua is with the University of California.

ABSTRACT

Heterogeneous networks, HetNets, are known to enhance the bandwidth efficiency and throughput of wireless networks by more effectively utilizing the network resources. However, the higher density of users and access points in HetNets introduces significant inter-user interference that needs to be mitigated through complex and sophisticated interference cancellation schemes. Moreover, due to significant channel attenuation and the presence of hardware impairments, e.g. phase noise and amplifier nonlinearities, the vast bandwidth in the millimeter-wave band has not been fully utilized to date. In order to enable the development of multi-Gigabit per second wireless networks, we introduce a novel millimeter-wave HetNet paradigm, termed *hybrid HetNet*, which exploits the vast bandwidth and propagation characteristics in the 60 GHz and 70–80 GHz bands to reduce the impact of interference in HetNets. Simulation results are presented to illustrate the performance advantage of hybrid HetNets with respect to traditional networks. Next, two specific transceiver structures that enable hand-offs from the 60 GHz band, i.e. the V-band to the 70–80 GHz band, i.e. the E-band, and vice versa are proposed. Finally, the practical and regulatory challenges for establishing a hybrid HetNet are outlined.

INTRODUCTION

The current generation of wireless standards, i.e. IEEE 802.11ac, can support data rates of up to 1.6 Gigabit per second (Gbps) employing high-order modulations and the multiple-input and multiple-output technology. In comparison, the latest wired Ethernet protocol can easily support data rates of up to 100 Gbps, i.e. more than 60 times faster. Thus, to ensure that wireless networking continues to be a viable alternative to wired networks, there is an urgent need for the development of multi-Gbps wireless links.

One approach for meeting the above need has been to more effectively share the network resources through the widely accepted notion of heterogeneous networks (HetNets) (Fig. 1). By employing smaller and more specialized cells, such networks can more efficiently meet the users' needs and improve the overall throughput of cellular networks [1]. However, the close vicinity of many users and cellular base station (BSs) along with the interference among these devices in HetNets, have introduced new challenges to the design of communication systems. Although many algorithms and approaches have been proposed for interference management and alignment in HetNets [2], these schemes are mainly complex in nature and may not be suitable for cost and power sensitive wireless applications.

Another approach for meeting the increasing demands for faster data relay connectivity is to take advantage of the vast bandwidth in the millimeter-wave band. It is noteworthy that the commissioned bandwidth by the Federal Communication Commission (FCC) in the 60 GHz and 70–80 GHz bands alone is 50 times the available bandwidth in today's cellular networks [3]. Thus, the millimeter-wave spectrum provides a great potential for meeting the tremendous demand for affordable and ultra high-speed wireless links. Current research has shown that point-to-point systems in the 70–80 GHz or the E-band can indeed support significantly higher date rates than the systems using the microwave band (frequencies below 30 GHz) [4]. Moreover, new wireless networking standards, e.g. IEEE 802.11ad and IEEE 802.15.3c, are designed to support multi Giga bit per second (Gbps) wireless networks in the 60 GHz or the V-band. In spite of these positive developments, millimeter-wave systems are significantly less bandwidth efficient than their counterparts in the microwave band [5–8].

As shown in Fig. 2, compared to the 60 GHz band, the atmospheric absorption in the 70–80 GHz spectrum is much lower, i.e. approximately 16 dB lower. Moreover, the FCC regulations allow for higher transmission power in the E-band compared to the V-band, i.e. maximum transmit power of 0.5 W and 3 W for the V-band and E-band, respectively [8]. Accordingly, due to the large attenuation factor and high antenna directivity, the 7 GHz of bandwidth in the 60 GHz band can be used to establish ultra high-speed wireless links without causing significant interference to neighboring devices and networks. These characteristics, which are beneficial from an interference management perspective, can also limit the ability of V-band systems to meet the quality of service (QoS) and throughput requirements of users in wireless or cellular networks.

0163-6804/15/$25.00 © 2015 IEEE

Using the aforementioned characteristics of the V-band and E-band, in this work we present a novel hybrid millimeter-wave based HetNet paradigm that enhances the bandwidth efficiency of millimeter-wave systems while reducing interference. More specifically, a hybrid HetNet makes use of the large bandwidth in the V-band to establish short range ultra high-speed point-to-point links. Due to the strong radio signal attenuation and high antenna directivity in the V-band, such links will not be a significant source of interference. Moreover, to circumvent the shortcomings of the V-band and to meet the QoS requirements of the network, we propose to apply the E-band spectrum to establish the longer-range links and to interconnect the HetNet BSs. For example, the links between the macrocell and picocell BSs and users can operate in the E-band, while the V-band can be used by femtocell BSs. Our simulation results illustrate the significant advantages of utilizing both the V-band and E-band in hybrid HetNet topologies. Next, new transceiver structures that allow the seamless handover from the V-band to the E-band and vice versa are presented, and their advantages and disadvantages are discussed. Finally, the challenges of implementing a hybrid HetNet structure in the millimeter-wave band, due to regulatory limitations, are presented and preliminary solutions are proposed.

The remainder of this article is organized as follows. The potential of utilizing the V-band or/and E-band in hybrid HetNets is presented. We focus on the transceiver structures and handoff issues in hybrid HetNets. We outline the main regulatory challenges for developing and implementing a Hybrid HetNet structure and present preliminary solutions to these challenges.

Potential of Hybrid Hetnets

The main challenges for the wide deployment of next generation wireless networks in the V-band can be summarized as [4, 7, 8]: 1) high path loss; 2) significant signal attenuation due to shadowing; and 3) limited transmit power. There are a number of solutions that are proposed for tackling these issues. These range from the deployment of highly directional antennas [7, 9–11] to the application of multi-beam directional arrays that radiate in multiple directions [12, 13]. The former approach requires the application of tracking algorithms that estimate the location of a user and adjust the beam pattern of the antenna mechanically or through signal processing approaches. Hence, this approach can be complex to implement and its performance is highly dependent on the accuracy of the tracking algorithm, which can vary substantially based on the propagation environment. By transmitting in multiple directions, the second approach seeks to mitigate the shadowing issue in the V-band, since it is anticipated that at least one of the beams or its reflections will reach the receiver. However, none of these schemes can address the limited operational range of V-band systems due to the significant channel attenuation and oxygen absorption.

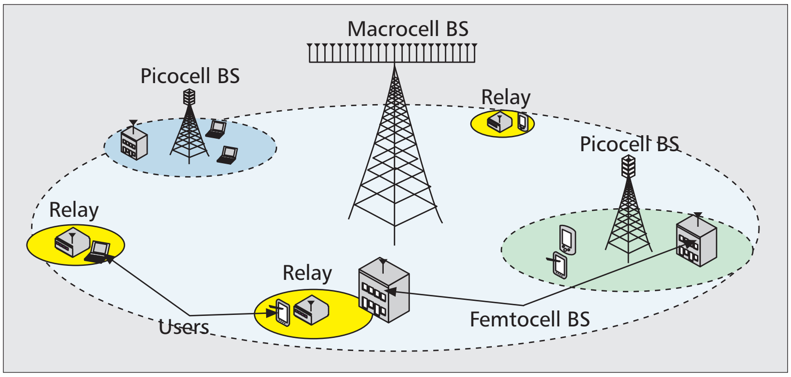


Figure 1. A HetNet with a macrocell BS and multiple supporting picocell BSs, femtocells, and relays.

Low SNR and low/high density	Medium-high SNR and high density	Medium-high SNR and low density
E-band	V-band	Both V- and E-bands

Table 1. Hybrid hetnet allocation of the V- and E-band spectra.

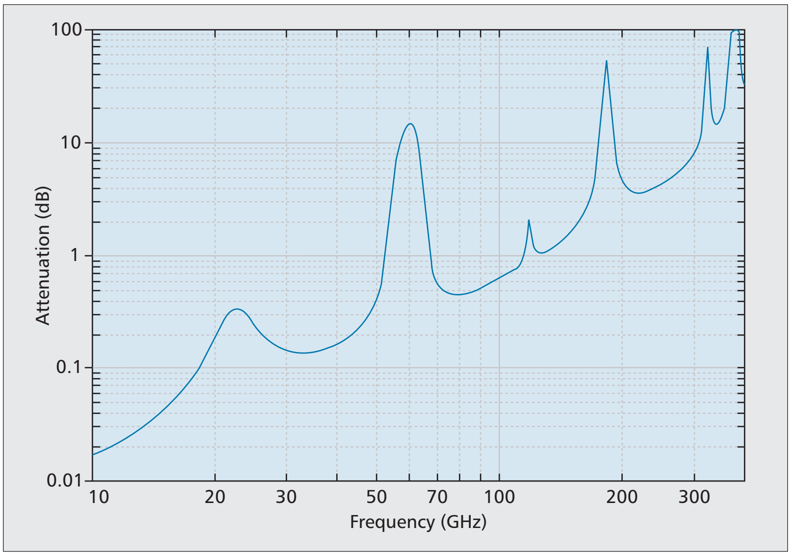


Figure 2. Atmospheric attenuation vs. operating frequency [8].

Unlike the V-band, the available spectrum in the E-band does not suffer from significant channel attenuation. As such, the 10 GHz of bandwidth in the E-band can be utilized to establish links with longer operational ranges [8]. Although these characteristics are advantageous and can address the shortcomings of the V-band, they can result in significant interference in densely deployed HetNets. Thus, we propose to concurrently use the bandwidths in the V-band and the E-band to achieve higher data rates, while reducing interference. Table 1 summarizes the framework for utilizing each or both bands in a hybrid HetNet configuration based on the network density and link signal-to-noise ratio (SNR). According to Table I, a hybrid HetNet topology can use the V-band in densely deployed networks. In this scenario, the high channel attenuation in the V-band ensures that the overall level of interference is reduced in the network. In addition, to maintain the QoS in low signal-to-noise ratio (SNR) scenarios, the E-band spectrum can be used instead. The lower channel attenuation and higher transmission power in the E-band can be used to increase the SNR. Finally, when extremely high throughputs are needed and the interference levels are low, both bands can be utilized.

In order to demonstrate the potential of a hybrid HetNet configuration, in Fig. 3 we compare the throughput of a point-to-point wireless link, while considering systems operating in the V-band, the E-band, and both V-band and E-band simultaneously. In this scenario the effects of channel attenuation, human shadowing, and interference from neighboring devices on the throughput of the link are also taken into account. Two tiers are taken into consideration. The network is assumed to consist of a macrocell and a femtocell BS and 10 users. The x-axis indicates the average distance among the users and their respective BSs. Our numerical results in Fig. 3 demonstrate the advantage of a hybrid HetNet compared to a HetNet that is operating in either the V-band or E-band. As shown in Fig. 3, as the link distance increases, the throughput of an "E-band system" is significantly higher than that of a "V-band system" due to a lower channel attenuation factor and higher maximum regulated transmission power. However, Fig. 3 shows that when taking the effect of interference into account, as the network density increases, a V-band wireless system can outperform its E-band counterpart. This can be attributed to the higher attenuation in the V-band, which reduces the overall inter-user interference in the network. As such, the results in Fig. 3 show that a HetNet that can take advantage of *either* the V-band or E-band can support wireless links over longer distances while reducing interference in densely deployed networks. Moreover, we observe that the overall performance of a wireless network can be further improved if the transceivers used in the hybrid HetNet can *simultaneously* transmit information over both the V-band and E-band spectra. This result is denoted by "Hybrid system using E- and V-band spectra" in Fig. 3.

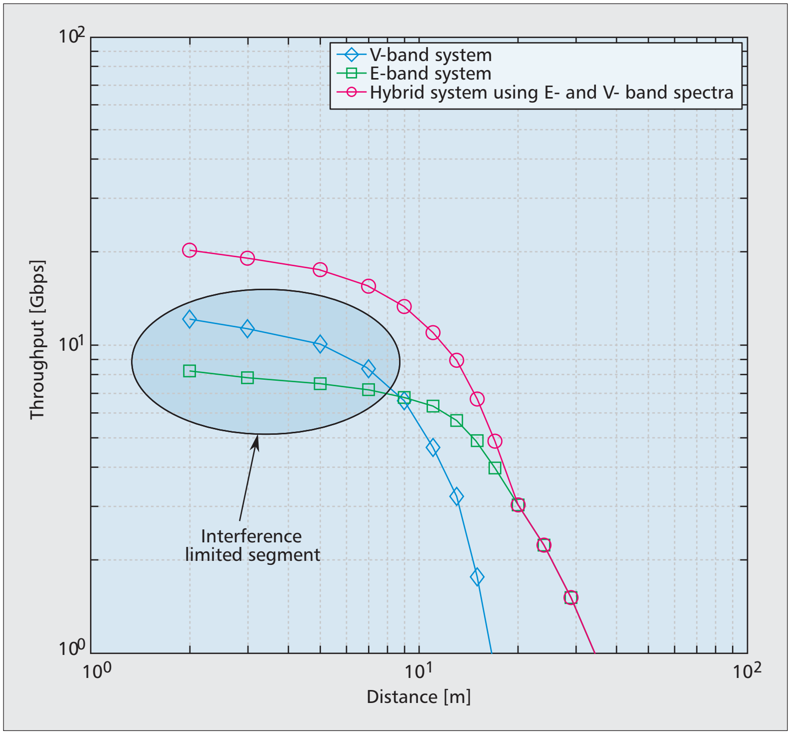


Figure 3. Throughput of a HetNet system when operating in either the V- or E-band, reduces the overall inter-user interference by utilizing both spectra (combined antenna gain of 30 dBi, bandwidth 5 GHz for both the V- and E-band, and 10-dB human shadowing).

Based on the results in Fig. 3, the precursor for transitioning or handing-off from one band to another can be due to the following reasons.

- In the strong interference regime, the access point or BS serving the wireless or cellular network, respectively, allocates the V-band spectrum to users with high SNRs, while assigning the E-band spectrum to users at lower SNR values. For example, the femtocells within a hybrid HetNet may use the V-band spectrum to avoid interfering with neighboring devices that are communicating with the macrocell or picocell BSs.
- In the low SNR scenario, to maintain the QoS, the access point allocates the E-band spectrum to users that have a link SNR below a certain threshold. For example, the E-band can be used to establish longer-range links that connect users at the cell edges to the macrocell BS.
- In the high SNR and low interference scenario, to meet the higher throughput requirements of a user, the access points may utilize the spectrum in both the V-band and E-band to meet the users' demands. This scenario can be considered in the context of transmitting audio/visual signals to an entrainment center. Such high-speed links can also be used to interconnect BSs.

Transceiver Structures for Handover Between V-Band and E-Band

Since the process of transitioning from the V-band to the E-band in hybrid HetNets is directly influenced by the transceiver design, we first propose two specific transceiver structures that support communication links in both the V-band and/or E-band spectra. The transceiver design in Fig. 4 allows for signal transmission in either the V-band or E-band, while the transceiver structure in Fig. 5 supports simultaneous transmission in both bands. The advantages/disadvantages of each design can be summarized as follows.

- Since the transceiver design in Fig. 4 uses a single band at any given time, it is possible to use a single voltage controlled oscillator (VCO) and amplifier at the radio frequency (RF) end. Moreover, as shown in Fig. 4, this transceiver structure does not require the deployment of a multiplexing block for transmission of the information bits in both the V-band and E-band. Therefore, the transceiver design in Fig. 4 is less complex and less costly to implement when compared to the design in Fig. 5.
- In general, designing an accurate and cost effective VCO that can operate over a very large set of frequencies is a rather challenging task. This design constraint, combined with the fact that the transceiver in Fig. 4 uses a single oscillator for 60 GHz, 70 GHz, and 80 GHz band trans-

mission, may make this transceiver more susceptible to the negative impact of disturbances, such as oscillator phase noise, compared to the transceiver in Fig. 5.

- One of the main challenges in the area of millimeter-wave communications is the design of RF amplifiers that can efficiently operate in this band. Keeping this in mind, we note that the transceiver design in Fig. 4 requires the application of an amplifier that can operate over a span of 30 GHz. Therefore, considering today's technology, it is anticipated that the transceiver in Fig. 4 may be more severely affected by amplifier nonlinearities and limited transmit power. On the other hand, the design in Fig. 5 may deploy standard amplifiers that are designed for either the V-band or E-band. This greatly simplifies the design process and may reduce the impact of impairments such as amplifier nonlinearities.
- The transceiver design in Fig. 5 can support higher data rates than the approach in Fig. 4, since it can use the spectra in both the V-band and E-band simultaneously.

Seamless handover between the V-band and E-band communication is essential for the effective operation of hybrid HetNets. Although very sophisticated handover schemes are available for current microwave based cellular networks [14, 15], these schemes may not be straightforwardly applied in the case of millimeter-wave systems, where the handovers take place over a much wider range of frequencies. For example, since the center frequencies for the V-band and E-band are separated by as much as 10 GHz, the transceivers intended for use in hybrid HetNets may require the application of two separate oscillators for each band. As such, compared to the microwave band, the handover between millimeter-wave frequencies is expected to be more severely affected by impairments, such as frequency offset and oscillator phase noise. This renders the process of achieving carrier synchronization during the handover process between the V-band and E-band to be far more challenging than microwave-based cellular networks. In addition, the handover schemes proposed for current cellular networks are designed to support transition from one cellular BS to another [14, 15], while in hybrid HetNets the transition from the V-band to E-band communication takes place among the same set of transceivers. This specific characteristic can be applied to reduce the overhead and complexity associated with handovers in hybrid HetNets. For instance, in the transceiver design in Fig. 5, while handing off from one band to another, the current link can serve as a feedback link to ease the process of channel estimation and synchronization during the handoff process. Thus, further research is needed to develop a comprehensive approach for seamless transmission from the V-band to the E-band and vice versa in hybrid HetNet configurations.

Regulatory Issues for the Development of Hybrid HetNets

Although the previous sections have demonstrated the advantages of using a hybrid HetNet structure, there are some regulatory issues that need to be addressed to make the deployment of such networks possible. In this section we briefly summarize these issues and provide solutions to each specific case.

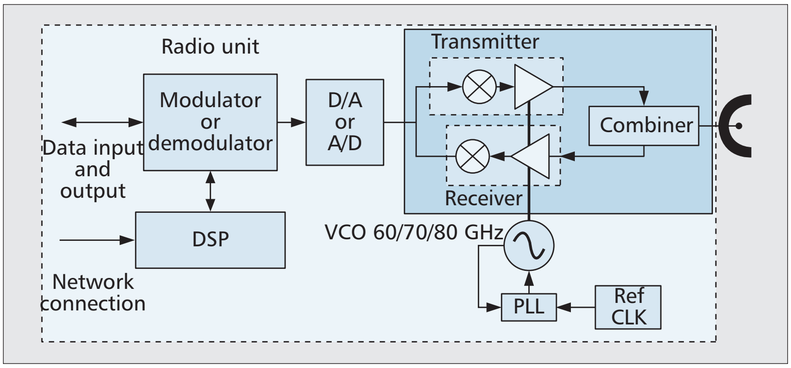


Figure 4. Transceiver structure for utilizing either the V-band or E-band.

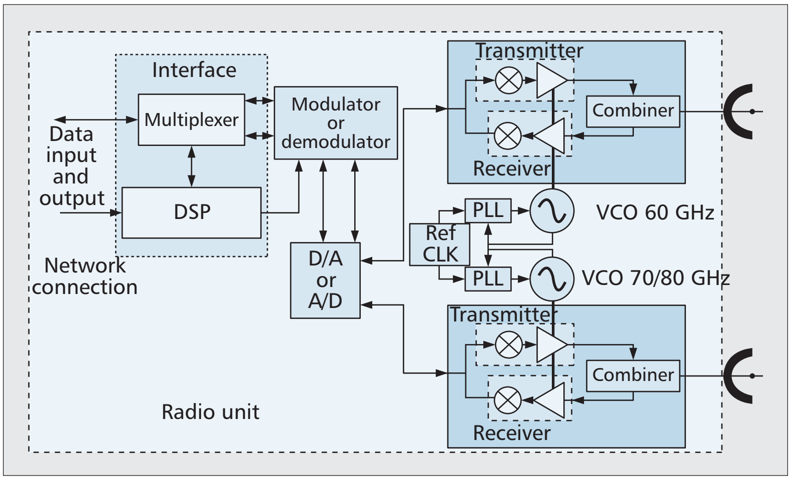


Figure 5. Transceiver structure for utilizing both the V-band and E-band spectra simultaneously (PLL, DSP, and CLK stand for phase locked loop, digital signal processing, and clock, respectively).

The V-band and E-band spectra are regulated or are being considered for regulation for the deployment of communication systems by most countries and regions in the world. More specifically, the V-band spectrum has been regulated for use in the United States, European Union, China, Canada, Korea, and many other countries around the world. Moreover, the E-band spectrum is regulated both in the United States and Europe and is being considered for the deployment of wireless communication systems in China, Canada, and many other countries. Although these regulations somewhat vary from country to country, the regulations proposed by the FCC represent a good sample of what is proposed or being proposed for these bands. We now summarize the FCC regulations regarding the use of both bands in Table 2.

Based on the regulatory requirements presented in Table 2, there are two specific issues that come to mind regarding the implementation of a hybrid HetNet structure.

- The spectrum in the E-band is licensed and may not be as readily available as the V-band for use in wireless and cellular networks. This presents new challenges for the deployment of hybrid HetNets by different cellular providers.

	V-band	E-band
Frequency range	57–64 GHz	71–76, 81–86, 92–95 GHz
Licensing	Unlicensed	Licensed
Maximum transmit power	27 dBm	35 dBm
Minimum antenna gain	Not applicable	43 dBi

Table 2. FCC regulatory rules for utilizing the V- and E-bands [8].

However, the FCC has adopted a unique licensing approach for spectrum allocation in the E-band, where the links can be quickly and economically registered over the Internet. It is also anticipated that, as the E-band is used more heavily due to spectrum needs, the licensing regulations on this band may be further relaxed.

- The high antenna gain required for the use of the E-band spectrum may render the application of this spectrum impractical in small and portable devices. Therefore, based on the current regulations, the E-band may be mainly utilized to carry the backhaul between access points and BSs, which in any case are long-distance links. However, as outlined above, it is anticipated that as the need for bandwidth and higher throughputs in wireless networks continue to grow, the large bandwidth in the E-band will also be made available for use via less restrictive regulations.

Summary

In order to reduce the impact of interference in HetNets and more effectively utilize the available spectrum in the millimeter-wave band, this article proposed a novel hybrid HetNet paradigm. In a hybrid HetNet, the V-band is used to establish short range and ultra high-speed point-to-point links, e.g. within femtocells. Due to significant channel attenuation in the 60 GHz band or the V-band, these short range links do not result in significant inter-user interference. On the other hand, the 70/80 GHz band or the E-band is used to establish the links with longer ranges, e.g. to interconnect cellular BSs. The E-band can support such links, since compared to the V-band, channel attenuation in the E-band is much smaller and the regulated maximum transmit power is much greater. Our simulation results demonstrated that by employing the characteristics of both bands, a hybrid HetNet configuration can greatly enhance the throughput of millimeter-wave networks. Moreover, two competing transceiver structures for hybrid HetNets were proposed and their advantages and disadvantages, from a practical standpoint, were discussed. We also outlined the regulatory challenges that need to be addressed to make a hybrid HetNet paradigm viable and provided preliminary solutions for each issue. Finally, although in this article the concept of HetNets has been discussed from a physical layer point-of-view, the successful deployment of such networks requires some modification to the higher layers. For example, the use of both the V-band and E-band may be determined by the higher layers based on the throughput needs of the user, while the physical layer facilitates the utilization of each band. Detailing these changes is beyond the scope of this article and is subject of future research.

References

[1] J. G. Andrews, “Seven Ways That HetNets are a Cellular Paradigm Shift,” *IEEE Commun. Mag.*, vol. 51, no. 3, Mar. 2013, pp. 136–44.

[2] A. Barbieri *et al.*, “LTE Femtocells: System Design and Performance Analysis,” *IEEE JSAC*, vol. 30, no. 3, June 2012, pp. 586–94.

[3] FCC, “Code of Federal Regulations, Title 47 — Telecommunications, Part 101: Fixed Microwave Services,” 2009.

[4] A. Georgiadis *et al.*, Eds., *Microwave and Millimeter Wave Circuits and Systems: Emerging Design, Technologies and Applications*, John Wiley and Sons Ltd, 2013.

[5] H. Mehrpouyan *et al.*, “Improving the Bandwidth Efficiency in E-band Communication Systems,” *IEEE Commun. Mag.*, vol. 53, no. 3, Mar. 2014, pp. 121–28.

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

[7] K.-C. Huang and Z. Wang, “Millimeter Wave Communication Systems,” IEEE Series on *Digital & Mobile Communication*, 2011.

[8] J. Wells, *Multi-Gigabit Microwave and Millimeter-Wave Wireless Communications*, Artech House, 2010.

[9] S. K. Yong and C. C. Chong, “An Overview of Multigigabit Wireless Through Millimeter Wave Technology: Potentials and Technical Challenges,” *EURASIP J. Wireless Commun. Network*, vol. 2007, 2007.

[10] A. Artemenko *et al.*, “High-Data-Rate Millimeter-Wave Radios,” *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, Apr. 2013, pp. 1665–71.

[11] R. C. Hansen, *Phased Array Antennas*, Wiley-Interscience, New York, 1998.

[12] H. Lee and H. Lee, “A Compact Dielectric Rod-Loaded Conical Horn Antenna for Millimeter-Wave Applications,” *Proc. Global Symposium Millimeter Waves (GSMM)*, May 2012, pp. 182–85.

[13] K. Huang and D. J. Edwards, “60 GHz Multi-Beam Antenna Array for Gigabit Wireless Communication Networks,” *IEEE Trans. Antennas Propag.*, vol. 54, no. 12, Dec. 2006, p. 3912–3914.

[14] R. Kim *et al.*, “Advanced Handover Schemes in IMT-Advanced Systems,” *IEEE Commun. Mag.*, vol. 48, no. 8, Aug. 2010, pp. 78–85.

[15] E. Dahlman, S. Parkvall, and J. Skold, *LTE/LTE-Advanced for Mobile Broadband*, Academic Press, 2011.

Biographies

Hani Mehrpouyan [S’05, M’10] (hani.mehr@ieee.org) received his B.Sc. honors degree in computer engineering from Simon Fraser University, Canada in 2004 and the Ph.D. degree in electrical engineering from Queen’s University, Canada in 2010. From 2010–2012 he was a post-doc at the Department of Signal and Systems at Chalmers University of Technology, where he lead the MIMO aspects of the microwave backhauling for next generation wireless networks project. He was also a visiting scholar at the University of Luxembourg in 2012, where he was involved in research related to interference cancellation for next generation satellite communication links. Since August 2012 he has been an assistant professor in the Department of Computer and Electrical Engineering at California State University, Bakersfield. For more information go to www.mehrpouyan.info.

Michail Matthaiou [S’05, M’08, SM’13] (m.matthaiou@qub.ac.uk) was born in Thessaloniki, Greece in 1981. He obtained the diploma degree (five years) in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece in 2004. He then received the M.Sc. (with distinction) in communication systems and signal processing from the University of Bristol, U.K., and Ph.D. degrees from the University of Edinburgh, U.K. in 2005 and 2008, respectively. From September 2008 through May 2010 he was with the Institute for Circuit Theory and Signal Pro-

cessing, Munich University of Technology (TUM), Germany, working as a postdoctoral research associate. He is currently a senior lecturer at Queen's University Belfast, U.K., and also holds an adjunct assistant professor position at Chalmers University of Technology, Sweden. His research interests span signal processing for wireless communications, massive MIMO, hardware-constrained communications, and performance analysis of fading channels. He currently serves as an associate editor for the *IEEE Transactions on Communications* and *IEEE Communications Letters*, and he was the lead guest editor of the special issue on "Large-Scale Multiple Antenna Wireless Systems" of the *IEEE Journal on Selected Areas in Communications*. He is an associate member of the IEEE Signal Processing Society SPCOM and SAM technical committees.

RUI WANG (liouxingrui@gmail.com) (M'14) received the B.S. degree from Anhui Normal University, Wuhu, China, in 2006, and the M.S. degree from Shanghai University, Shanghai, China, in 2009, and the Ph.D. degree from Shanghai Jiao Tong University, China, in 2013, all in electronic engineering. From August 2012 to February of 2013 he was a visiting Ph.D. student in the Department of Electrical Engineering at the University of California, Riverside. From October 2013 to October 2014 he was with the Institute of Network Coding, The Chinese University of Hong Kong, as a postdoctoral research associate. Since October 2014 he has been with the College of Electronics and Information Engineering, Tongji University as an assistant professor. His research interests include wireless cooperative communications, MIMO technique, network coding, and OFDM etc.

GEORGE K. KARAGIANNIDIS [SM'03, F'14] (geokarag@ieee.org) was born in Pithagorion, Samos Island, Greece. He received the university diploma (five years) and Ph.D. degree, both in electrical and computer engineering, from the University of Patras, in 1987 and 1999, respectively. From 2000 to 2004 he was a senior researcher at the Institute for Space Applications and Remote Sensing, National Observatory of Athens, Greece. In June 2004 he joined the faculty of Aristotle University of Thessaloniki, Greece, where he is currently professor and director of the Digital Telecommunications Systems and Networks Laboratory. In January 2014, he joined Khalifa University, UAE, where is currently Professor in the Electrical & Computer Engineering Dept. and Coordinator of the ICT Cluster. His research interests are in the broad area of digital communications systems, with emphasis on communications theory, energy efficient MIMO and cooperative communications, cognitive radio, smart grid, and optical wireless communications. He is coauthor of the book *Advanced Wireless Communications Systems* (Cambridge Publications, 2012). He has been a member of Technical Program Committees for several IEEE conferences such as ICC, GLOBECOM, VTC, etc. In the past he was editor for fading channels and diversity of the *IEEE Transactions on Communications*, senior editor of *IEEE Communications Letters*, and editor of the *EURASIP Journal of Wireless Communications and Networks*. He was lead guest editor of the special issue on "Optical Wireless Communications" of the *IEEE Journal on Selected Areas in Communications*, and guest editor of the special issue on "Large-Scale Multiple Antenna Wireless Systems". Since January 2012 he has been the editor-in-chief of *IEEE Communications Letters*.

YINGBO HUA [S'86, M'88, SM'92, F'02] (yhua@ee.ucr.edu) received a B.S. degree in 1982 from Southeast University, Nanjing, China, and a Ph.D. degree in 1988 from Syracuse University, Syracuse, NY. He held a faculty position with the University of Melbourne, Australia, where he was promoted to the rank of reader and associate professor in 1996. He was a visiting professor with Hong Kong University of Science and Technology (1999-2000), and a consultant with Microsoft Research, WA (summer 2000). Since 2001 he has been a professor with the University of California at Riverside.

Dr. Hua has served as editor, guest editor, a member of the Editorial Board and/or member of the Steering Committee for *IEEE Transactions on Signal Processing*, *IEEE Signal Processing Letters*, *EURASIP Signal Processing*, *IEEE Signal Processing Magazine*, *IEEE Journal on Selected Areas in Communications*, and *IEEE Wireless Communications Letters*. He has been a member of IEEE Signal Processing Society's Technical Committees for Underwater Acoustic Signal Processing, Sensor Array and Multichannel Signal Processing, and Signal Processing for Communication and Networking. He has served on the Technical and/or Organizing Committees for more than 50 international conferences and workshops. He has authored/coauthored more than 300 articles and coedited three volumes of books, with more than 8000 citations, in the fields of sensing, signal processing and communications. He is a Fellow of IEEE and AAAS.

10 Gb/s HetSNets with Millimeter-Wave Communications: Access and Networking – Challenges and Protocols

Kan Zheng, Long Zhao, Jie Mei, Mischa Dohler, Wei Xiang, and Yuexing Peng

Kan Zheng, Long Zhao, Jie Mei, and Yuexing Peng are with Beijing University of Posts and Telecommunications.

Mischa Dohler is with King's College London.

Wei Xiang is with the University of Southern Queensland.

ABSTRACT

Heterogeneous and small cell networks (HetSNets) increase spectral efficiency and throughput via hierarchical deployments. In order to meet the increasing requirements in capacity for future 5G wireless networks, millimeter-wave (mmWave) communications with unprecedented spectral resources have been suggested for 5G HetSNets. While the mmWave physical layer is well understood, major challenges remain for its effective and efficient implementation in HetSNets from an access and networking point of view. Toward this end, we introduce a novel but 3GPP backwards-compatible frame structure, based on time-division duplex, which facilitates both high-capacity access and backhaul links. We then discuss networking issues arising from the multihop nature of the mmWave backhauling mesh. Finally, system-level simulations evaluate the performance of HetSNets with mmWave communications and corroborate the possibility of having capacities of tens of gigabits per second in emerging 5G systems.

INTRODUCTION

The traffic load of wireless communications networks with various quality of service (QoS) requirements has recently been increasing rapidly due to the widespread use of mobile Internet applications by smart terminals. This trend continues, requiring emerging wireless networks to be designed to meet these requirements. The fifth generation (5G) cellular communications system is expected to be standardized around 2020 [1]. The overarching goal of 5G is to achieve 10 to 100 times higher user data rates such that in dense urban environments the typical user data rate will range from 1 to 10 Gb/s, while supporting 10 to 100 times more connected devices. Moreover, to facilitate the vision of the tactile Internet, end-to-end latency will need to be less than 5 ms so as to provide ultra-fast application response times.

As a result, novel physical, access, and network layer technologies are required to realize such ambitious goals. Generally speaking, there are several means to improve network performance. The first one is to increase the available bandwidth (e.g., carrier aggregation or cognitive radios). The second way is to increase geographic spectrum reusability through, say, device-to-device (D2D) communication and small cell techniques [2, 3]. The third but not least is to improve spectral efficiency, such as (massive) multiple-input multiple-output (MIMO) and non-orthogonal multiple access (NOMA) techniques. However, even though some of these techniques are able to boost performance substantially, there is no clear roadmap on how to achieve the so far defined 5G performance targets. Breakthrough technologies are thus needed in the near future.

Low frequency bands have been almost used up in recent years, and it is difficult to find sufficient frequency bands in the microwave range for 5G. By contrast, there are still a large number of unused frequency bands in the millimeter-wave (mmWave) bands, which may be of potential use to mobile cellular communications given recent advances in hardware and electronic components. At present, mmWave communications have already had numerous indoor and outdoor applications. It is well suited for not only in-home applications like audio/video transmission, desktop connections, and portable devices, but also outdoor point-to-point applications. MmWave communications also play an increasingly significant role as the wireless backhaul for outdoor small cells, thanks to its low costs, quick deployment, and flexibility. Apart from these benefits, the evolved NodeB (eNB) in a small cell may provide high system throughput through mmWave communications. The propagation characteristics of mmWave communications are characterized by a high level of oxygen absorption and rain attenuation, especially in outdoor environments. This limits the range and cell coverage of mmWave radio as opposed to microwave radio.

On the other hand, mmWave communications systems facilitate integration with the increasingly popular massive MIMO technology [4]. This is because the wavelength of mmWave radio is small enough so that the physical size of

0163-6804/15/$25.00 © 2015 IEEE

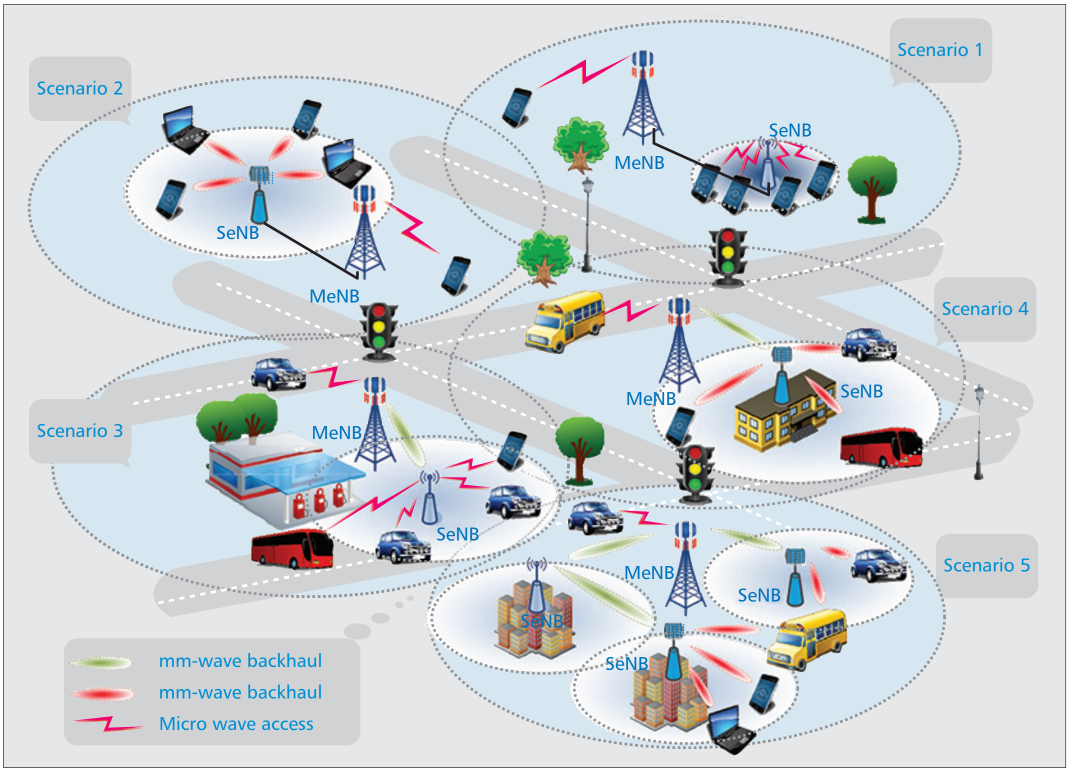


Figure 1. Illustration of typical deployment scenarios with both mmWave and microwave systems.

Heterogeneous and small cell networks (HetSNets) with mmWave communications will play a very important role in future 5G cellular networks. However, there are many problems related to the implementation of HetSNets with mmWave communications capabilities.

a massive antenna array can be greatly reduced for easy deployment at small-cell eNBs. Apart from improving spectral efficiency, massive MIMO is also an effective technique to compensate for the severe propagation loss of mmWave radio [5]. Recently, beamforming/precoding techniques with massive MIMO in mmWave have been studied widely in IEEE standards, such as IEEE 802.15.3c (TG3c) for wireless personal area networks (WPANs), IEEE 802.11ad (TGad), and the Wireless Gigabit Alliance (WiGig). Meanwhile, massive MIMO with 3D beamforming is becoming a much discussed topic in Release 12 of the 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) [6].

Therefore, heterogeneous and small cell networks (HetSNets) with mmWave communications will play a very important role in future 5G cellular networks. However, there are many problems related to the implementation of HetSNets with mmWave communications capabilities. While there is ample physical (PHY)-layer material available, such as [7], to the best of the authors' knowledge there have been few reported studies in the literature addressing the communications challenges from an access and networking point of view [8]. The scope of this article is thus to study the interplay of mmWave and microwave communications in HetSNets from the access and networking points of view.

Toward this end, the article is organized as follows. We briefly discuss typical deployment scenarios for HetSNets with mmWave communications. Given the scenario requirements, we then design a 3GPP-LTE-compliant time-division duplex (TDD) frame structure capable of supporting multihop transmission via a wireless backhaul. We discuss medium access and networking related challenges and solutions for HetSNets with mmWave communications. We then present and discuss performance results of previously introduced protocols. Finally, conclusions are drawn.

Deployment of mmWave Communications in HetSNets

A HetSNet typically consists of multiple types of radio access nodes in a 3GPP LTE network, for example, a macrocell eNB (MeNB) and multiple small cell eNBs (SeNBs) such as pico, femto, and relay eNBs. In such a network, each SeNB combines its backhaul data with that received from other nodes in the network before forwarding it to the MeNB. The SeNBs are supposed to be separated by short distances (e.g., 100~200 m), which helps mitigate severe propagation losses. Coverage within the small cells (i.e., user access) may also be provided by mmWave radio, reducing the level of interference experienced on the sub-3 GHz frequency bands used for traditional cellular communications.

Typical Deployment Scenarios

With the introduction of mmWave communications in the HetSNet, there are several potential deployment scenarios with mmWave communications being used for backhaul and/or user access links. Some typical scenarios are illustrated in Fig. 1 and discussed in more detail below.

Scenario 1 (Baseline): Traditionally, an SeNB is connected to its donor MeNB through a wired backhaul such as an optical fiber. All user equipments (UEs) are served by either the MeNB or

SeNBs on the microwave band. Under this scenario, interference coordination schemes have to be carefully designed so as to avoid interference between the MeNB and SeNBs.

Scenario 2: Compared to scenario 1, the UEs communicating with the MeNB work on the microwave band, whereas the UEs served by the SeNBs use mmWave radio. The SeNBs are connected to the MeNB via a wired backhaul. No interference coordination is needed between the MeNB and SeNBs under this scenario. However, the UEs should support dual bands for smooth handover between the MeNB and SeNBs, thus increasing the costs of the UEs.

Scenario 3: In lieu of a wired backhaul, the mmWave band is employed for backhaul transmission between the MeNB and SeNBs. For the sake of implementation, only single-hop is permitted for backhaul transmission with mmWave radio. Through such a deployment, only network facilities need to be upgraded, while the UEs remain unchanged. This is helpful for quick deployment of SeNBs. Similar to scenario 1, advanced interference coordination schemes are necessary between the MeNB and SeNBs.

Scenario 4: MmWave communications are adopted for single-hop wireless backhaul for the SeNBs in scenario 4. Moreover, the SeNBs serve the UEs in a small cell via mmWave radio, which can significantly increase network capacity thanks to the tremendous bandwidth offered by the mmWave band.

Scenario 5: Increasing geographic spectrum reusability is another means to improve network capacity, resulting in dense small cell deployment. Then multihop wireless backhaul is a good way to connect dense SeNBs with the MeNB. In this scenario, the SeNBs can cooperate with one another and communicate with their donor MeNB via mmWave radio. Also, the access links between the SeNBs and their served UEs work on the mmWave band. This scenario is much more flexible and can provide high capacity, in which most key techniques for HetSNets with mmWave communications may be applicable.

Since Scenario 1 is the baseline, and Scenarios 2–4 are generally a subset of Scenario 5, we subsequently study the more general Scenario 5 from an access and networking point of view. Based on that scenario as well as the unique characteristics of mmWave communications, we subsequently design an improved 3GPP LTE backward-compatible access and backhaul frame structure.

Frame Structure

The propagation characteristics of mmWave communication bands are fairly different from those of microwave communications, such as the Doppler frequency shift and multipath delay. Therefore, the orthogonal frequency-division multiplexing (OFDM) parameters for microwave communications in 3GPP LTE systems are *not* applicable to mmWave communications without modifications. On the other hand, much larger frequency bandwidths are available on mmWave communication bands, which means the bandwidth per subcarrier needs to be enlarged in order to keep unchanged the size and complexity of the fast Fourier transform (FFT). Moreover, backward compatibility with 3GPP LTE systems must be considered when introducing mmWave communications into HetSNets. To facilitate the utmost efficient system design, a new frame structure has to be designed for mmWave communications coexisting with microwave communications in HetSNets.

Backward compatibility with 3GPP LTE systems must be considered when introducing mmWave communications into HetSNets. To facilitate the utmost efficient system design, a new frame structure has to be designed for mmWave communications coexisting with microwave communications in HetSNets.

Assume that 28 GHz band with 1 GHz in bandwidth is used for mmWave communications due to its availability and popularity [9]. Measurement results show that the delay spread is not severe if the transmitter-receiver (Tx-Rx) separation is less 200 m at 28 GHz in the urban environment. In the line of sight (LOS) case, there are too few multipaths to determine the root mean squared (RMS) delay spread, while in the case of non-line-of-sight (NLOS), most measured multipath components have RMS delay spreads below 0.2 μs. Thus, OFDM-based mmWave wireless systems need to tolerate an RMS delay of 0.2 μs or less. The guard interval for mmWave communications is expected to be much larger than the RMS delay in order to alleviate intersymbol interference (ISI). Besides this, backward compatibility with microwave communications has to be taken into consideration.

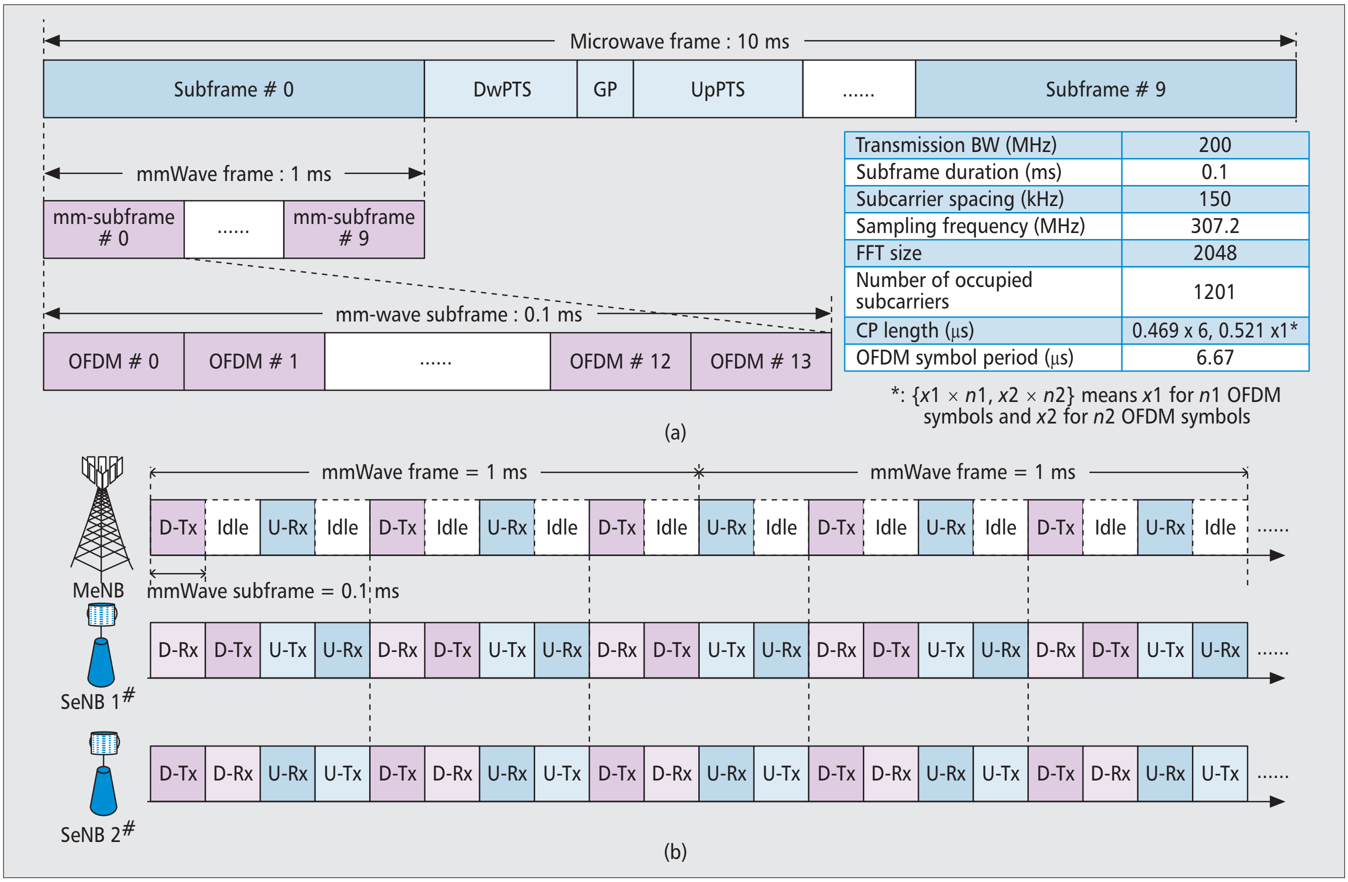
Therefore, a guard period of 0.469 μs is selected, because this value is not only two times more than the measured RMS delay but also one tenth[1] of the guard period in microwave communications. A symbol period of 66.67 μs is selected in 3GPP LTE, which is 66.67/4.69 times larger than the guard period. Correspondingly, a useful symbol duration of 6.667 μs can be chosen for mmWave communications and the subcarrier spacing is computed as the inverse of the OFDM useful symbol duration (i.e., 150 kHz). Then the basic mmWave frame has 10 subframes, each of which is of 0.1 ms duration and consists of 14 OFDM symbols. The frame structure and main parameters are depicted in Fig. 2a as one possible TDD solution to HetSNets with mmWave communications.

When wireless backhaul is adopted in the HetSNet with mmWave communications, the mmWave subframe configuration is able to support multihop transmission. Each mmWave subframe can be used for single-hop transmission. Figure 2b illustrates an example configuration for multihop transmission. In the first mmWave subframe, the MeNB transmits backhaul signals to the nearby SeNB 1# in the first-hop transmission, and SeNB 2# far from the MeNB sends information to its small cell UEs (SUEs) simultaneously. Subsequently, the MeNBs remain silent in the second mmWave subframe, and SeNB 1# transmits signals to SeNB 2# and their SUEs, taking advantage of spatial multiplexing. The uplink procedure is similar to that of the downlink. Moreover, in order to reduce the Rx/Tx switch, several mmWave subframes can be grouped together to provide radio resources for a single hop.

MAC and Networking Design Challenges

The use of highly directional beamforming raises a number of new challenges in the network design. We focus only on the medium access

[1] Note that this approach is the inverse of a recent machine-type communications (MTC) approach, where the channel is downsampled by a factor of 10. In the future, we envisage legacy 3GPP LTE/LTE-A systems run at the sampling rate, whereas low-complexity MTC devices run on a tenth of the sampling rate, and ultra-high-capacity mmWave systems at a 10 times higher sampling rate.



Transmission BW (MHz)	200
Subframe duration (ms)	0.1
Subcarrier spacing (kHz)	150
Sampling frequency (MHz)	307.2
FFT size	2048
Number of occupied subcarriers	1201
CP length (μs)	0.469 x 6, 0.521 x1*
OFDM symbol period (μs)	6.67

*: {$x1 \times n1$, $x2 \times n2$} means $x1$ for $n1$ OFDM symbols and $x2$ for $n2$ OFDM symbols

Figure 2. Illustration of the proposed frame structure: a) modified TDD frame structure for mmWave communications; b) mmWave subframe configuration for multihop communications.

control (MAC) and networking layers in this article. Thanks to the dense deployment of small cells using narrow beams with large gains on mmWave communication bands, multihop transmission is made possible in HetSNets. Routing schemes with a specific metric under certain delay constraints and reasonable control overheads become very important. Different from interference coordination in microwave communications, mainly in either the time or frequency domain, spatial interference coordination receives more attention in mmWave communications, and is thus discussed in this section. Moreover, due to the limited coverage of mmWave communications, control channels are implemented on microwave communication bands for connectivity and mobility management, and data channels via mmWave communication bands.

Routing in Multihop HetSNets

In dense urban areas, a large number of small cells may be deployed closely in HetSNets. There are backhaul connections among the SeNBs via mmWave radio. The backhaul between the SeNBs and their donor MeNB can be via either mmWave radio or wired links, depending on the deployment scenarios. In such a network, the nodes can cooperate with each other, providing improved reliability, enhanced coverage, and reduced equipment costs. Compared with mobile ad hoc networks, route recovery and energy efficiency are not major concerns for multihop HetSNets due to limited mobility and the existence of power supply at the eNBs. Moreover, there is almost no interference between mmWave signals due to the narrow beamform with large gains. In other words, the channels for connecting the SeNBs can be regarded as orthogonal, so no channel assignment is needed. However, each channel has its own propagation characteristics, and network topology is determined by the deployment scenario. Therefore, how to design an efficient routing protocol becomes one of the key challenges in multihop HetSNets.

When designing a routing scheme, one needs to taken into account several objectives, such as increasing system throughput, decreasing end-to-end delay, and achieving a good load balance. Thus, much attention is paid to obtaining a reliable quantitative routing metric, which can link several factors (e.g., system throughput, end-to-end delay, and connectivity) with the quality of the routing scheme. In general, we can define such a metric as a function of N influence factors I_j (i.e., $\gamma = f(I_1, I_2, \cdots, I_N)$). For example, there are two QoS factors that mainly affect the routing performance: the end-to-end delay τ and the packet loss rate η of the path. Then the routing metric can be defined as a linear weighted sum of the two influence factors, $\gamma = (1 - \beta)\tau + \beta\eta$, which determines the importance of the delay relative to the loss rate, and where β is a weighting coefficient.

Most existing routing algorithms are based on

Since the signal strength deteriorates rapidly in the mmWave band, more than one eNB can transmit simultaneously if they are not very close to each other. Thus, all the SeNBs are grouped into clusters according to the distances not only between themselves but also to their donor MeNB.

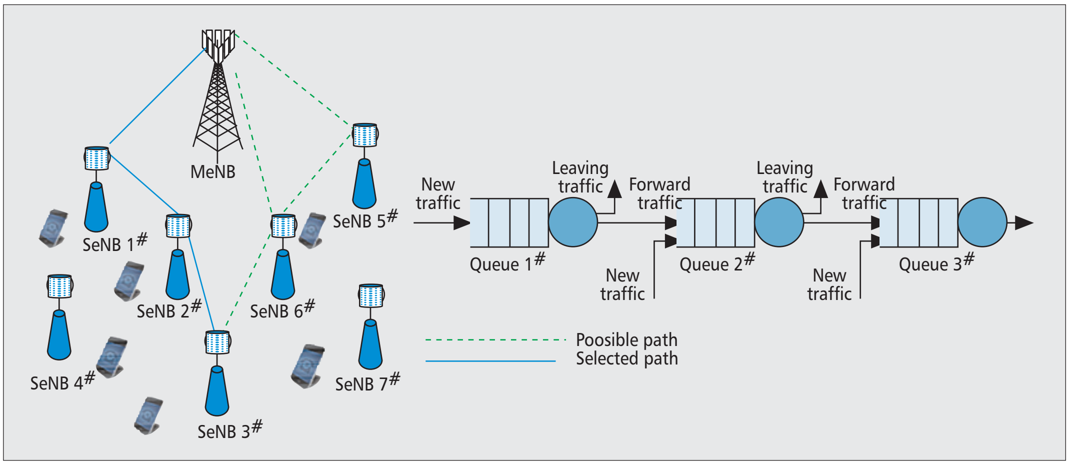


Figure 3. A routing example in multihop HetSNets.

the minimum hop count, which is of no interest for this article. Instead, through guaranteeing some end-to-end QoS requirements, QoS routing algorithms with specific routing metrics become much more promising in future broadband wireless communications networks. Queue theory is a good theoretical tool to solve QoS-related problems. In a multihop wireless network, not only the traffic arrival process but also the variation of the wireless channel can be modeled by a completed tandem queuing framework [10]. The queuing dynamics of nodes in a given path determines the end-to-end QoS metrics including the delay and loss rate, which have to be measured in a timely fashion in order to enable the routing algorithm to adapt to the changes of the network in time. The end-to-end delay is the sum of the delays that any packet experiences in all the queues and links along its routing path. Packets may be lost due to buffer overflow in one of the queues in tandem.

As shown in Fig. 2, a number of consecutive time slots form a fixed-sized time frame, where the time slots in each time frame are periodically allocated to some transmission links. Representing one hop, each link may be allocated time slots in each time frame with different spatial orthogonal channels formed by mmWave beamforming signals. Figure 3 presents an example of a possible path from an SeNB to the MeNB and its corresponding tandem queue. Data traffic entering each queue may come from different connections. Traffic from connections other than the considered connection may come and leave the tandem system in any queue. In other words, in each queue of the tandem system (except queue 1), either new or forwarding traffic arrives while the buffered traffic is leaving. There are K single queues along the route. The arrival traffic to each queue is from the previous queue of the tandem system and from other connections traversing the corresponding link (except for queue 1). Given the allocated radio resources for each link along the tandem system, we can determine the service rate. Thus, if the arrival probability for the aggregate traffic is determined, the queuing performance measures for each queue in the tandem system can be calculated.

Assume that all the end-to-end performance measures for a general tandem system of queues with an arbitrary number of hops can be found. Then the tandem queuing model can be applied to find a path of connections from a source node to its desired destination node so that the end-to-end QoS requirements for the connection can be satisfied. Given an optimization objective (e.g., minimizing the end-to-end delay), the best feasible route can be found by exhaustively searching all possible routes. However, this approach results in large signaling overhead and prohibitively high computational complexity. Therefore, investigating new routing schemes with the tandem queuing model in consideration of feasible implementation complexity becomes one of the challenges in studying the multihop HetSNet with mmWave communications.

Moreover, the route information of HetSNets can be updated periodically in the control channel to keep track of the changes in topology and traffic load of the network. To reduce the control overhead and delay, hierarchical routing schemes may be applied in HetSNets. Several neighboring SeNBs can be grouped together to form a cluster. Then routing can be performed either intra-cluster or inter-cluster, depending on service requirements.

Access Control and Interference Coordination

Thanks to the narrow beam in massive MIMO with mmWave radio, spectral resources can easily be reused spatially. The interference among various links and cells in the HetSNet with mmWave communications becomes much simpler than that with only microwave radio. However, since multihop transmission is allowed, and each eNB works in the half-duplex mode (still), the interference between the downlink and uplink has to first be avoided through interference coordination among the eNBs in the time domain. Since the signal strength deteriorates rapidly in the mmWave band, more than one eNB can transmit simultaneously if they are not very close to each other. Thus, all the SeNBs are grouped into clusters according to the distances not only between themselves but also to their donor MeNB. The eNBs in the same cluster can-

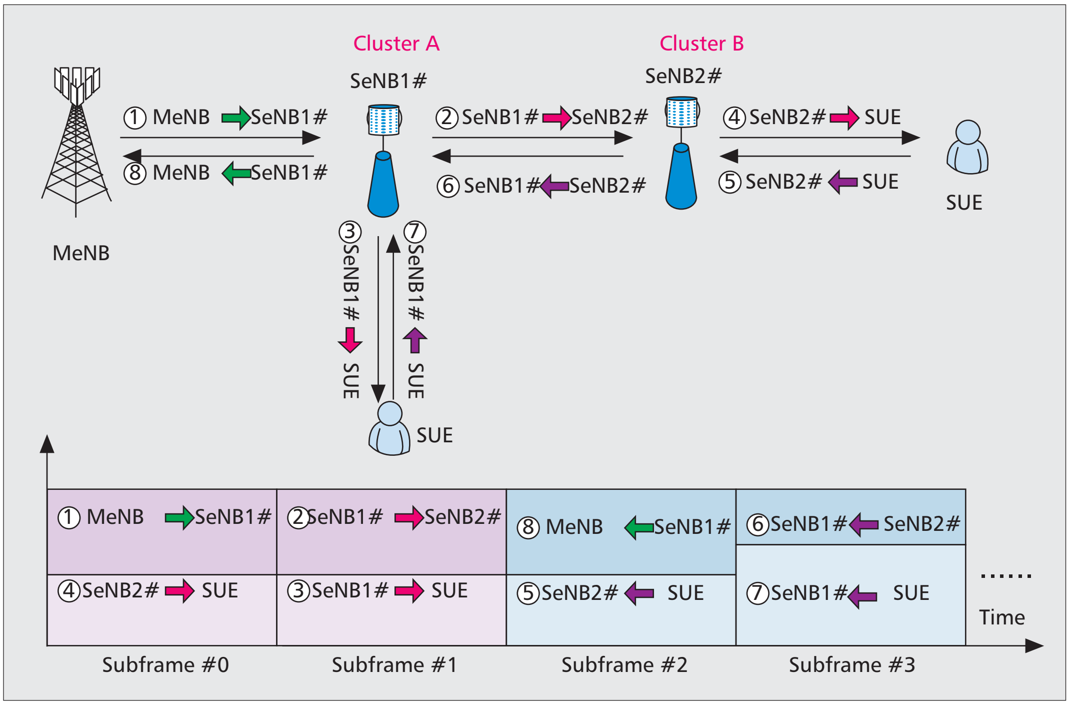


Figure 4. Illustration of radio resources allocation in multihop HetSNets.

Although dense small cells can facilitate geographic spectrum reusability, their small cell size may cause overly frequent handover. Consequently, it is imperative to introduce mechanisms for mobility management in the HetSNet with mmWave radio.

not transmit simultaneously, whereas those in different clusters can if the distance between the clusters is large enough. As illustrated in Fig. 4, the MeNB and SeNB 1# belong to Cluster A, while SeNB 2# is in Cluster B because it is a little far from the MeNB. Assuming that the configuration in Fig. 2b is adopted in the system, the MeNB establishes downlink backhaul transmission with SeNB 1# in subframe 0, while SeNB 2# also serves its SUEs in this subframe. In subframe 1, SeNB 1# transmits signals to both its SUEs and SeNB 2# distinguished by narrow beams, while the MeNB remains silent. A similar procedure is performed in subframes 2 and 3. Thus, by properly grouping all the SeNBs, interference can be avoided between the downlink and uplink in the HetSNet with mmWave communications.

Since the interference between the eNBs can be nearly eliminated through beamforming, neither intracell nor intercell coordination is necessary on the downlink. In other words, each eNB can reuse all the spectral resources. As shown in Fig. 4, due to the limitation of SUEs, distinguishing uplink signals cannot rely only on the spatial domain. In subframe 2, there is almost no interference between the access link from SUE to SeNB 2# and the backhaul link from SeNB 1# to MeNB due to spatial reuse. However, the signals from SUEs without beamforming arrive at SeNB 1# simultaneously in subframe 3, as well as the wireless backhaul signal sent from SeNB 2#. They may interfere with each other if no action is taken. A feasible way to avoid possible interference is to make the signals orthogonal in the frequency domain. Since the beamformer with large gains can still be used in the backhaul link, much less spectral resources are needed for the backhaul link of SeNB 2# to SeNB 1# than the access link from SUEs to SeNB 1#. Moreover, if there are more than one SeNB in cluster A, the uplink interference between the access links of the small cells can be coordinated by specific soft-frequency reuse schemes.

MmWave Softcell Concept

Although dense small cells can facilitate geographic spectrum reusability, their small cell size may cause overly frequent handover. Consequently, it is imperative to introduce mechanisms for mobility management in the HetSNet with mmWave radio. Motivated by the soft cell concept [6], control channels are provided by the MeNB with microwave communications, while high date rate services are supported by the SeNBs with mmWave communications in the small cells simultaneously (i.e., separated C-plane/U-plane configuration). Also, the MeNB mainly ensures wide-area coverage so as to maintain good connectivity and mobility. One of the carriers on the microwave bands is selected as the anchor carrier, in which the system and control signaling information are sent by the MeNB. When the control channels in the anchor carrier work normally, the SeNBs do not send cell-specific signals/channels, such as primary/secondary synchronization signals (PSS/SSS) and cell-specific reference signals (CRS). However, the resources for the control channels are reserved in the small cells and become active in the event of emergency, which is managed by the MeNB under the master-slave mode. Usually, the radio resource control (RRC) connection procedures such as channel establishment and release between the SUEs and SeNBs are controlled by the MeNB through the anchor carrier.

One of the most important benefits of the separation between C-plane and U-plane lies in its robustness to handover. There is no handover when a UE moves from one small cell to another within the coverage of the MeNB. Meanwhile, the requirements on the RRC messages can be relaxed because of the low handover probability between the macrocells. Also, energy efficiency can be improved by the separated C-plane/U-plane configuration in a massive deployment of small cells. For example, some SeNBs can be turned off when there are no serving SUEs. However, this requires new features of the UEs. Dual transmit channels and MAC entities in both the microwave and mmWave carriers have to be supported in the UEs. Then all the carriers can be measured and discovered rapidly.

Performance and Discussions

System-level simulations for different cases of Scenario 5 have been carried out to evaluate the system performance of HetSNet with mmWave communications, where we evaluate the downlink throughput. Detailed simulation parameters, including the channel model and system assumptions, are summarized in Table 1, most of which are defined in the 3GPP specifications [11]. All the UEs are evenly distributed in circular areas around their donor eNB. For simplicity, an ideal tractable piecewise-linear array pattern is used in the simulations [12]. This pattern is a good approximation to the practical radio pattern. Assuming broadside transmission (i.e., maximum gain at the azimuth angle of 0), the normalized array gain at an azimuth angle θ is given by

$$g(\theta)=\begin{cases} 1, & \text{for } \theta<\theta_1 \\ 1-\dfrac{|\theta|-\theta_1}{2(\theta_{3dB}-\theta_1)}, & \text{for } \theta_1<|\theta|\leq\theta_2 \\ \text{FBR}, & \text{for } \theta_2<\theta\leq\pi, \end{cases} \quad (1)$$

where FBR denotes the front-to-back ratio, and $\theta_{3\mathrm{dB}}$ is the half-power beamwidth. The variables $\theta_{3\mathrm{dB}}$, θ_1, and θ_2 control the array manifold shape, half-power beamwidth, and FBR, respectively.

Parameters		Values
Carrier frequency/bandwidth (microwave)		2 GHz/ 10 MHz
Carrier frequency /bandwidth (mmWave)		28 GHz/ 200 MHz
MeNB inter-site distance (ISD)		500 m
Cellular layout		Hexagonal grid/ 19 sites/3 sectors per site
Small cell deployment		2, 4, 10 SeNBs per sector
UE density		60 UEs per site
Total MeNB Tx power		46 dBm
Total SeNB Tx power		37 dBm
MeNB antenna configuration		2
SeNB antenna configuration		Linear array
UE antenna configuration		1
Noise figure at UE		9 dB
Thermal noise density		–174 dBm/Hz
Penetration loss (mmWave)		100 dB
Shadowing standard deviation		8 dB microwave/ 12 dB mmWave
Path loss	Microwave	128.1 + 37.6log(*R*), *R* in km
	mmWave	157.4 + 32log(*R*), *R* in km [5]

Table 1. Parameters assumption in HetSNets with mmWave communication.

As shown in Fig. 5a, we consider densely populated areas where there are multiple SeNBs deployed in each sector that is at most three hops away from the MeNB. Placement of the SeNBs may significantly affect the performance of the HetSNets. Usually, the SeNBs are placed as close as possible to the cell edge, while a qualified backhaul link can be maintained by a high-gain beam. Also, enough distance between each SeNB should be maintained in order to avoid excessive interference among one another. Then, in our simulations for Cases 1 and 2 (i.e., 2 or 4 SeNBs/sector, respectively, as illustrated in Fig. 5a), the SeNBs are placed on the circles around the centers of the hexagons with a radius of 1/9 intercell distance (ISD) in consideration of cell edge priority. In Case 3 (i.e., 10 SeNBs per sector), the SeNBs are placed in two two-tier circles, with radii of 1/10 ISD and 1/5 ISD for the inner and outer circles, respectively. Each SeNB can independently schedule its connected UEs according to a certain channel-aware scheduling algorithm on the mmWave frequency band. A small number of SUEs may occupy the entire mmWave frequency band, resulting in very high data rate transmission.

Figure 5b presents the spatial characteristics of the signal-to-interference-plus-noise ratio (SINR) of the HetSNet under different cases. Different colors represent different throughput values. For example, the area marked in red has a higher SINR than that marked in green or blue. It is clear that there are more areas with the red color in Cases 2 and 3 than in Case 1. The SINR performance is thus dramatically increased by deploying more SeNBs. However, such an improvement is more obvious in Case 2 than in Case 3. When a large number of SeNBs with full frequency reuse among cells are deployed (e.g., 10), the close distance between the SeNBs may cause interference and decrease the SINR even with narrow beams on the mmWave frequency band.

The performances including the average SUE throughput and aggregate throughput (TP) of small cells in the networks under different cases are compared in Table 2. In Case 1, only 21.7 percent of all the UEs are served by the SeNBs with an average throughput per SUE of 333.5 Mb/s due to sparse deployment. With the increase of SeNBs deployed, more UEs choose to connect to the SeNBs with single or multihop

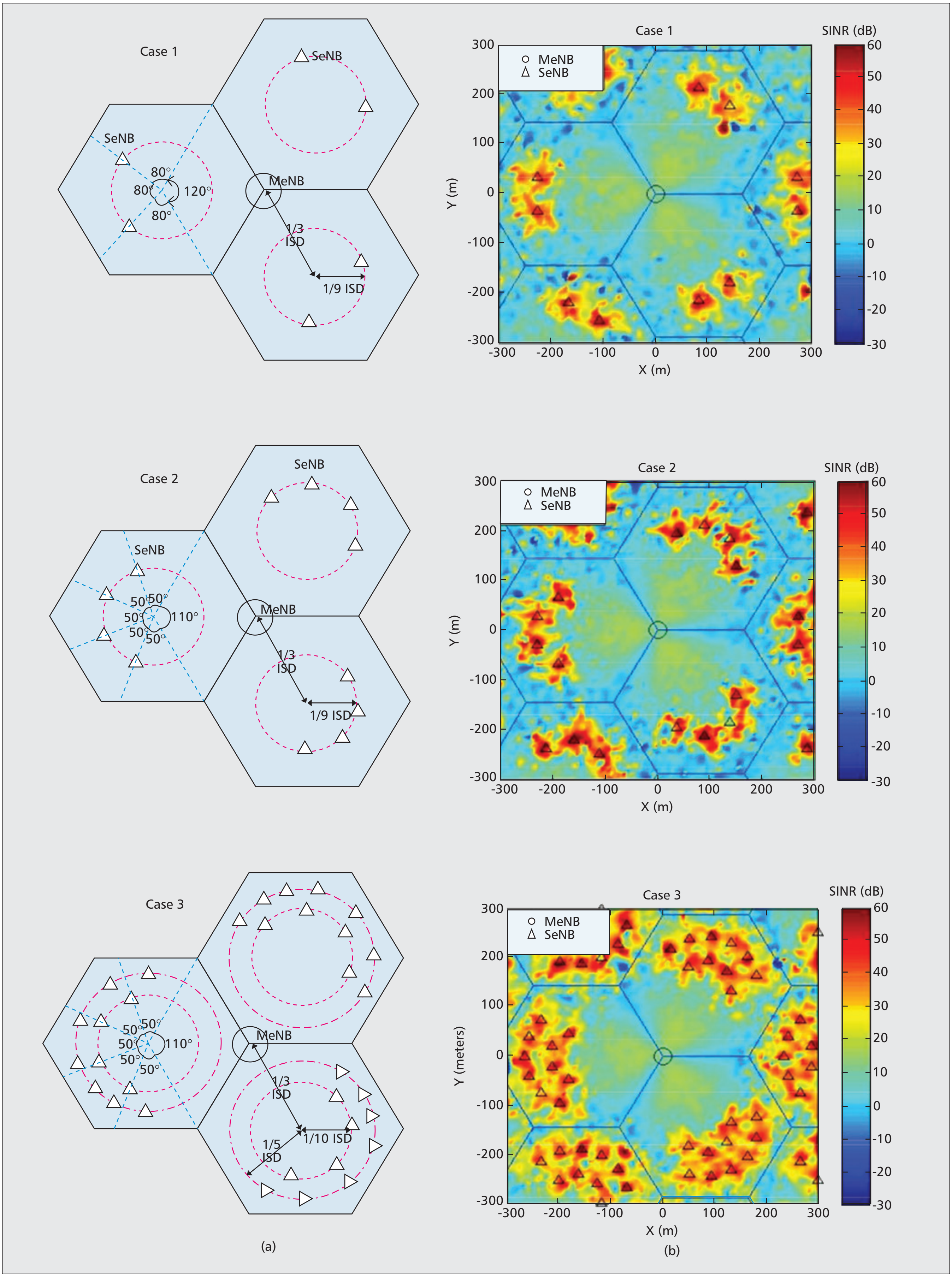


Figure 5. Illustration of deployment and SINR performances in HetSNets with mmWave communications: a) SeNB deployment; b) geometry of SINR performance.

With the development of new mmWave physical layer techniques, the fundamental knowledge on mmWave communications becomes more solid. Future research on HetSNets with mmWave communications will then be motivated by a tight coupling of the unique characteristics of mmWave communications and wireless heterogeneous networks.

	Average throughput per SUE (Mb/s)	Average number of SUEs per cell	Ratio of SUEs/all UEs	Total cell throughput (Gb/s)
Case 1	276.9	13	13/60 = 21.7%	3.64
Case 2	422.2	20	22/60 = 33.3%	8.26
Case 3	388.2	34	34/60 = 56.7%	12.98

Table 2. Performance comparisons in HetSNets under difference cases.

transmission (i.e., 33.3 and 56.7 percent of all the UEs in Cases 2 and 3, respectively), since the distance between SeNBs and UEs is shorter. Meanwhile, more small cells means that the same frequency resources can be more frequently reused with the aid of beamforming techniques. Therefore, not only the average throughput per SUE but also the total throughput rapidly increase. Moreover, in this way, more severe co-channel interference due to close distance arises when a larger frequency reuse factor is adopted. The average throughput per SUE slightly decreases in Case 3 compared to that of Case 2. However, because many more UEs may access the network through their nearby SeNBs in Case 3 than in Case 2, the total throughput performance continues to be improved when the number of SeNBs increase from 4 to 10 (i.e., 12.98 Gb/s), almost doubling that in Case 2.

Conclusions and Future Work

HetSNets with mmWave communications are critical to meeting the requirements of future 5G wireless communications. Due to the vast spectral resources of mmWave radios, users will be able to enjoy unprecedented services with almost wire-like user experience. We investigate the feasibility of mmWave communications in HetSNets under various deployment scenarios, and propose a new 3GPP LTE-A backward-compliant frame structure. This allows us to address several important challenges facing mmWave communications, and achieve an aggregated cell throughput of nearly 13 Gb/s, an order of magnitude more than the current best 5G system design [13].

With the development of new mmWave physical layer techniques, the fundamental knowledge on mmWave communications becomes more solid. Future research on HetSNets with mmWave communications will then be motivated by a tight coupling of the unique characteristics of mmWave communications and wireless heterogeneous networks. Specifically, a more in-depth study of decoupling the control and data channels is needed, which will lead to new handover, call admission control, and radio resource management protocols. Furthermore, mmWave offloading mechanisms ought to be studied from the load and queue points of view.

Acknowledgment

The work was supported in part by the China Natural Science Funding (61271183), the National High Technology Research and Development Program of China (2014AA01A705), the Fundamental Research Funds for the Central Universities (NO.2014ZD03-02), and the Program for New Century Excellent Talents in University (NCET-11-0600).

References

[1] "Scenarios, Requirements and KPIs for 5G Mobile and Wireless System," ICT-317669-METIS/D1.1, May 2013.
[2] L. Lei *et al.*, "Operator Controlled Device-to-Device Communications in LTE-Advanced Networks," *IEEE Wireless*, vol. 19, no. 3, June 2012, pp. 96–104.
[3] K. Zheng *et al.*, "Energy-Efficient Wireless In-Home: The Need for Interference-Controlled Femtocells," *IEEE Wireless Commun.*, vol. 18, no. 6, Dec. 2011, pp. 36–44.
[4] E. G. Larsson *et al.*, "Massive MIMO for Next Generation Wireless Systems," *IEEE Commun. Mag.*, vol.52, no. 2, Feb. 2014, pp. 186–95.
[5] Z. Pi and F. Khan, "A Millimeter-Wave Massive MIMO System for Next Generation Mobile Broadband," *Proc. IEEE ASILOMAR*, 2012, pp. 693–98.
[6] D. Astely *et al.*, "LTE Release 12 and Beyond," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013, pp. 154–60.
[7] S. Hur *et al.*, "Millimeter-Wave Beamforming for Wireless Backhaul and Access in Small Cell Networks," *IEEE Trans. Commun.*, vol. 61, no. 10, Oct. 2013, pp. 4391–03.
[8] J. Qiao *et al.*, "MAC-Layer Concurrent Beamforming Protocol for Indoor Millimeter Wave Networks," *IEEE Trans. Vehic. Tech.*, to appear
[9] T. S. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," *IEEE Access*, vol. 1, May 2013, pp. 335–49.
[10] L. Le and E. Hossain, "Tandem Queue Models with Applications to QoS Routing in Multihop Wireless Networks," *IEEE Trans. Mobile Computing*, vol. 7, no. 8, Aug. 2008, pp. 1025–40.
[11] 3GPP TR 36.814, V9.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA): further Advancements for E-UTRA Physical Layer Aspects," 2010.
[12] S. Akoum, O. E. Ayach, and R. W. Heath, "Coverage and Capacity in mmWave Cellular Systems," *Proc. IEEE ASILOMAR*, Nov., 2012, pp. 688–92.
[13] P. Blasco *et al.*, "Aggressive Joint Access & Backhaul Design for Distributed-Cognition 1 Gbps/Km2 System Architecture," WWIC 2010, 1–3 June 2010, Lulea, Sweden, invited paper.

Biographies

Kan Zheng [SM'09] (zkan@bupt.edu.cn) received his B.S., M.S. and Ph.D. degrees from Beijing University of Posts & Telecommunications (BUPT), China, in 1996, 2000, and 2005, respectively, where he is currently a professor. He worked as a senior researcher in companies including Siemens and Orange Labs R&D (Beijing), China. His current research interests lie in the field of wireless communications, with an emphasis on resource allocation in heterogeneous networks and M2M/V2V networks. He has published more than 200 papers in IEEE conferences and transactions.

Long Zhao is studying toward a Ph.D. at BUPT. His research interests include the next generation mobile communications, massive MIMO, and wireless green communications.

Jie Mei received his B.S. degree from Nanjing University of Posts &Telecommunications, China, in 2013. Since then, he has been working toward an M.S. degree at BUPT. His research interests include massive MIMO, millimeter-wave mobile communications, and heterogeneous networks.

MISCHA DOHLER [F '13] is a professor in wireless communications at King's College London, a member of the Board of Directors of Worldsensing, a Distinguished Lecturer of the IEEE, and Editor-in-Chief of *Transactions on Emerging Telecommunications Technologies*.He is a frequent keynote, panel, and tutorial speaker. He has pioneered several research fields, contributed to numerous wireless broadband and IoT/M2M standards, holds a dozen patents, organized and chaired numerous conferences, has more than 200 publications, and authored several books. He has a citation h-index of 38 (top 3 percent).

WEI XIANG [M'04, SM'10] received his B.Eng. and M.Eng. degrees, both in electronic engineering, from the University of Electronic Science and Technology of China, Chengdu, in 1997 and 2000, respectively, and his Ph.D. degree in telecommunications engineering from the University of South Australia, Adelaide, in 2004. Since January 2004, he has been with the School of Mechanical and Electrical Engineering, University of Southern Queensland, Toowoomba, Australia, where he currently holds a faculty post of associate professor. He was a co-recipient of the Best Paper Award at IEEE WCNC '11. He has been awarded several prestigious fellowship titles. His research interests are in the broad area of communications and information theory, particularly coding and signal processing for multimedia communications systems.

YUEXING PENG [M'08] is an associate professor in wireless communications at BUPT. His research interests include physical layer technologies and digital signal processing. He holds a dozen patents, and has published more than 50 papers and books.

Millimeter-Wave Multimedia Communications: Challenges, Methodology, and Applications

Dan Wu, Jinlong Wang, Yuming Cai, and Mohsen Guizani

Dan Wu, Jinlong Wang, and Yuming Cai are with PLA University of Science and Technology.

Mohsen Guizani is with Qatar University.

ABSTRACT

The worldwide opening of a massive amount of unlicensed millimeter-wave spectrum has triggered great interest in developing high-bit-rate multimedia services and applications. Specific challenges for mmWave communication design include large-scale attenuation, atmospheric absorption, phase noise, limited gain amplifiers, and so on. This article aims to define and evaluate important metrics to characterize multimedia QoS and jointly takes these technical challenges into account in the framework of mmWave. To this end, we design a QoS-aware multimedia scheduling scheme to achieve the trade-off between performance and complexity, in which accurate propagation analysis is carried out and suitable countermeasure techniques are pointed out to satisfy the QoS requirements. Moreover, potential multimedia applications are analyzed and possible solutions provided. Illustrative results indicate that the proposed multimedia scheduling scheme can perform efficiently in a practical mmWave communication system.

INTRODUCTION

Although more and more users prefer to enjoy high-quality low-latency multimedia applications, content providers are still limited to a carrier frequency spectrum ranging between 700 MHz and 2.6 GHz [1]. In fact, the rapid increase of mobile Internet and wireless services are posing unprecedented challenges for wireless telecommunication operators to overcome a global bandwidth shortage [2]. Millimeter-wave (mmWave) communication becomes a possible choice for the next generation broadband wireless communication technology (e.g., 5G). As such, a natural question arises in order to find solutions to the bandwidth shortage, which is how to develop multimedia applications in the environment of mmWave communication. The characteristics are briefly summarized below by comparing them to microwave ultra-wideband (UWB) technology.

First, UWB is a promising technology due to its unprecedented system bandwidth in the unlicensed band of 3.1–10.6 GHz. The low emission and impulsive nature of a UWB radio (at short distance) results in enhanced bit rate and communication security. Moreover, the UWB radio can operate easily with low energy cost and computational complexity. It is able to transmit high-bit-rate multimedia contents with some QoS guarantees. However, the most challenging issue for UWB is that international coordination regarding the operating spectrum is difficult to achieve among different countries. However, spectrum allocation is not an issue for mmWave [2]. In addition, mmWave is able to provide high data rates, and its spectrum allocation can be coordinated worldwide [3]. This is the basis of the popularity of mmWave.

The second characteristic is the high bit rate requirements. For the current UWB systems, they can at most provide a data rate of 480 MB/s, which is far less than the requirements of uncompressed video applications. For example, bit rates for high-definition TV usually exceed 2 GB/s. Although the emerging OFDM-UWB system is able to support 2 GB/s, the implementation complexity, power consumption, and network cost will increase exponentially compared to mmWave.

The third characteristic is the variation of the received signal attenuation at a given spectrum. For the OFDM-UWB systems, there are five band groups covering a frequency range from 3.1 to 10.6 GHz. According to the Friis propagation model [4], the propagation attenuation is inversely proportional to the square of a group center frequency given the same transmitted power. In contrast, because of relatively small changes in frequency, the coverage range does not dynamically change for mmWave.

Finally, the intersystem interference is another issue to be considered. The UWB band covers the 2.4 to 5 GHz unlicensed bands (for WiFi/802.11); thus, mutual interference is a major concern. In order to enable existing wireless systems to be implemented in different environments, regulatory authorities in different areas are working on their own requirements for UWB operation. It is impossible for a regional UWB radio to work in another region, but international harmonization for mmWave is possible (e.g., an IEEE standard group has been estab-

0163-6804/15/$25.00 © 2015 IEEE

lished [3]). This action constitutes the necessary condition for a wider use of mmWave applications.

In this work we study mmWave multimedia communications from the aspects of technical challenges, design methodology, and possible applications. In particular, an investigation of the mmWave design framework is provided, and the corresponding performance evaluation is carried out. The aim of this work is to design a unified and general methodology and criteria for mmWave communications, and to show how these technologies can be applied to multimedia applications in a general framework of mmWave. The rest of this work is organized as follows. The technical challenges are introduced. Then a general design methodology is provided, and possible multimedia applications are discussed. Numerical results and conclusions are then provided, respectively.

Technical Challenges

Although mmWave technologies promise a bright future for multimedia communications, current multimedia transmission protocols and technologies cannot be applied directly due to the technical challenges encountered in the wireless channels, multimedia coding, and transmission mode. The above technical challenges should be well analyzed and understood to design an appropriate communication system.

Wireless Channels

It is well known that a wireless channel attenuation is a measure of the average signal intensity or energy decay as a function of distance, frequency, and other location-specific parameters. Previous research work on the 60 GHz mmWave communications found that the wireless channel attenuation suffers from 20–40 dB loss in terms of signal energy, and additional 7–15.5 dB/km loss due to the atmospheric absorption phenomenon. Specifically, Guo *et al.* [3] compared the signal attenuation at 60 and 2.5 GHz, and drew the conclusion that the material attenuation at 60 GHz is usually much higher than that of 2.5 GHz. This effect, combined with the traditional wireless propagation rule [4], acts as a basic result for wireless channel conditions for mmWave communications.

In fact, the condition of the wireless channel at mmWave is much more complex than the aforementioned result for 60 GHz. In particular, different mmWave communications have various sensitivities to the transmission environment (e.g., glass, water, buildings). It is extremely difficult to capture a general channel attenuation relationship [4]. Instead, we can divide the transmission environment into different categories and provide an approximate relationship based on a body of experimental results. This kind of method is being widely used in mmWave communications today.

Moreover, it is important to note that multiple paths in the mmWave propagation channel generate multipath interference, which cannot be ignored in mmWave communications [3]. Furthermore, when the signal attenuation is combined with multipath interference, the path length differences are obvious due to surface reflections, object scattering, and heterogeneous propagations. In particular, surface reflections and object scattering are substantially reduced when the wavelength is close to the object size, and heterogeneous propagations depend on the ability to penetrate solid substances [3]. Another issue to be considered is the Doppler frequency, which is proportional to the transmitted signal frequency, and represents the maximum frequency difference between received and transmitted signals due to object mobility [2]. As a result, the higher the carrier frequency, the greater the Doppler effects. To avoid the Doppler effect in mmWave communications, we should describe the relationship between the Doppler effect and the varying channel at different carrier frequencies and transmission environments. Many experiments have been conducted to estimate this relationship, and a fundamental result is that the Doppler effect for mmWave communications is about 10 times higher than traditional wireless communications [4].

Multimedia Coding

To capture the transmitted and received multimedia content via mmWave communications, the operating systems of mmWave provide multimedia application programming interfaces for multimedia communications software to coordinate multimedia coding, transmission, and signal loss. In particular, the software possesses the following properties. It determines the coding rate at which the multimedia content suffers from the minimum information distortion. Similar to classical wireless multimedia transmissions, the coding rate plays an important role in quality of service/experience (QoS/QoE) guarantee, which achieves some sense of balance among system congestion, multimedia distortion, and communication throughput. In terms of mmWave communications, Yang *et al.* [7] suggested that at least 25 frames/s is recommended to provide a high-definition video service, 20 frames/s for CIF video format. Hence, this is the minimum coding rate for different multimedia services via mmWave communications.

As usual, to reduce the bandwidth requirements of mmWave, multimedia applications should be compressed before transmission. Current video coding standards, H.264 and MPEG-4, are both able to provide a satisfactory QoS with relatively low bit rates. In particular, the authors in [6] have designed a general multimedia service architecture that has received substantial attention recently. However, these standards cannot be applied directly in mmWave communications due to the following possible causes:

- The quantization step is dynamic and cannot be estimated precisely in a high attenuation environment.
- Motion estimation cannot be accurately implemented as the difference between adjacent blocks becomes smaller.
- Coding complexity is so huge that it cannot be operated online, in particular for energy-limited mobile devices.

Therefore, although the current state-of-the-art coding methods have become very mature with efficient compression and low latency, they

To capture the transmitted and received multimedia content via mmWave communications, the operating systems of mmWave provide multimedia application programming interfaces for multimedia communications software to coordinate multimedia coding, transmission, and signal loss.

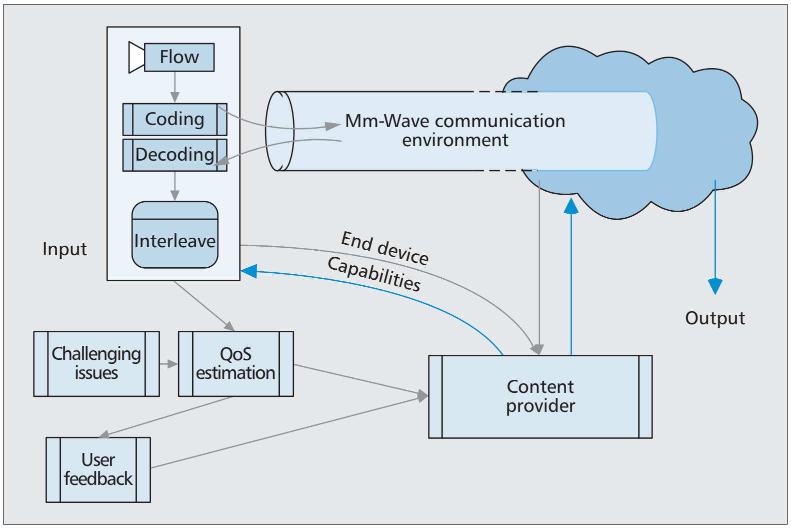


Figure 1. General millimeter-wave multimedia transmission framework.

need to be further revised for mmWave. Some important parameters, such as frame rate, group of picture size, quantization parameter, and resolution, should be reconsidered or re-examined under the new environment. A regression method is used to capture these parameters with different carrier frequencies, and establish the relationship using a relatively simple model (e.g., a linear or exponential model). Moreover, advanced scalable video coding is also a possible solution, but it inevitably yields increased computational costs and overall bit rates. As a result, more studies should focus on balancing the trade-off between the benefit and cost in a general mmWave communication framework.

Transmission Mode

The transmission mode for mmWave communications differs from device to device and from service to service. For simplicity, device-to-device transmission can be combined with mmWave for short-distance communications. However, in this case the energy consumption should increase dramatically since the received energy is huge compared to traditional wireless communications.

Note that the choice of a suitable transmission scheme is primarily dependent on the system objective, that is, one spends as few resources as possible with respect to time, frequency, and energy consumption. The expected resource in turn relies on the transmission distance and carrier frequency. In addition, as we know, multi-cell transmission reduces intercell interference by achieving higher spectrum efficiency. As a result, multi-cell transmission can be used in this case. However, it may require transmission in cells without users, thus making the multi-cell scheme inefficient as soon as there are several participating cells without any users. In contrast, a single-cell transmission needs few resources, but strongly depends on the reception conditions of users, and thus it is difficult to apply in mmWave communications.

Given the previous analysis, the determination of the exact transmission fashion is too costly on the transmission interface. Motivated by [4–6], for the mmWave communications framework, there are two possible mechanisms to decide which transmission scheme to use. The first one is the determination of the maximum transmission distance with at least one user for a given multimedia application with QoS, implicitly providing information on the transmission distance and thus a reasonable breakpoint between transmission schemes. The second one is the comparison between energy consumption needed for multi-cell transmission and single-cell transmission according to the feedback from the devices and the signal intensity.

Methodology

In this section we propose a general multimedia transmission methodology (shown in Fig. 1) for mmWave communications. It takes into account the multipath propagation, signal attenuation, signal interference, and communication noise. Moreover, it employs channel coding and communication diversity to counteract the negative effects of previous technical challenges.

System Architecture

As shown in Fig. 1, the content provider acts as the source of the transmission system with dynamic bitstream bit rates. These bitstreams are either continuous or intermittent multimedia flows due to the characteristics of the multimedia services and applications. In the first step, the crude bitstream is encoded by the channel coding to improve error tolerance; interleaving technology can be applied here. Subsequently, the output of channel coding sends it to the baseband processing block, the objective of which is to realize multiplexing functions for multiple points and link transmissions. In particular, in the multicast scenario multiplexing provides an estimated capacity as the maximum transmission rate. At the output of this step, to obtain a spectrally shaped baseband complex signal, an advanced signal transform method can be utilized to reduce the relationship between multimedia flows. After that, the modulated signal can be produced by a simple modulator for message transmissions.

To offset the frequency selectivity and explore the diversity of a broadband channel, orthogonal frequency-division multiplexing (OFDM) technology can be used by combining it with advanced source technologies. These united methods can avoid adaptive equalization and time-cost leanings, but results in a nonlinear carrier estimation with huge computational complexity. In addition, the frequency selectivity caused by the multipath channel is also counteracted in OFDM systems by splitting the content into multiple sub-band carriers. In this case each sub-band can satisfy flat fading conditions when the number of subbands is chosen appropriately. Furthermore, the modulation and demodulation parts can be operated by introducing inverse fast Fourier transform and Laplace transform, respectively.

In a multi-carrier mmWave communication system, each carrier suffers from different fading losses due to the frequency selectivity of the channel. A possible solution for studying the relationship between the frequency selectivity and diversity is to use error-resilient compressed transmission. As a result, in this work we combine the graph-based coding method with Reed-Solomon coding through a signature authentication. Another possible solution is to use the unicast mode, in which one should precisely select a subset of the carriers to each user in a dynamic and smart manner. The additional benefit of this method lies in taking the frequency diversity into account.

In some extreme cases (e.g., when the required bandwidth is far larger than the available resources) a traditional single-carrier method is more suitable. In this case multipath fading can be alleviated by point-to-point transmission. Moreover, channel coding can be introduced at the sender and receiver to combat the signal loss due to multipath fading. For example, in the downlink of a Long Term Evolution (LTE) communication system, the base station can easily select the most appropriate channels (e.g., list the channel conditions with priority) for point-to-point communication. Then some kind of channel coding can be embedded into the coded stream through achieving the trade-off between multimedia distortion and system congestion, as described in [8]. Similarly, different choices of modulation lead to different trade-offs between performance and complexity. For detailed information, readers are referred to [9].

Scheduling Scheme

For a dynamic mmWave multimedia communication system, each user reports the QoS requirement to each access control part, which decides whether the multimedia service is acceptable or not based on the results of the QoS estimation using the aforementioned method. In terms of a relatively static network topology with short transmission distance, there are more choices depending on the available energy, QoS requirements, and transmission delay. Usually, each user individually performs the distributed scheduling scheme [7, 8]) based on his/her local information. It should be noted that a user here denotes all users in the system, including the new ones accepted by the access control part. The whole scheduling procedure is illustrated in Fig. 2.

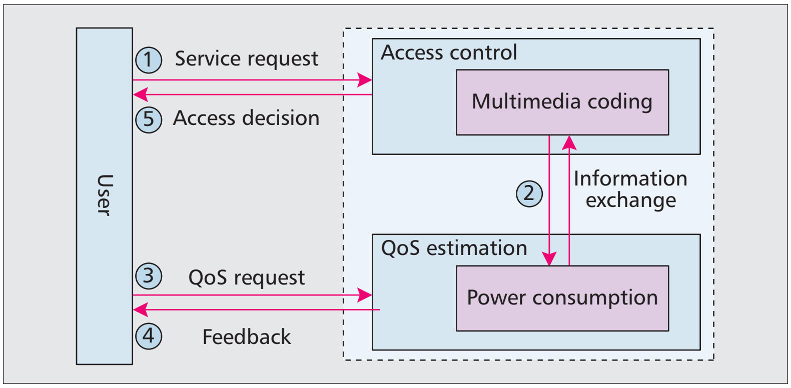


Figure 2. Dynamic scheduling for mmWave multimedia communication system.

Subsequently, each user passes the power consumption and multimedia coding file to the access control part, and this file is treated as the initial value of the multimedia application once it is accepted. After conducting the first-round QoS estimation, the access control part replies to each user with an instruction that indicates whether the desired QoS is achievable or not. The new user is accepted if it receives a positive reply; otherwise, each user has to perform higher transmitting power or shorter transmission distance. Then a new round of information exchange and judgement between the user and access control part is started. When all the possibilities are not allowed, the new user/application is rejected. Note that there are many existing attempts in the literature that focus on how to find all the possibilities faster. Here we refer readers to [8] for a detailed implementation. From the above procedure, we can roughly analyze the overhead of the proposed method as follows: user report costs overhead is $O(1)$; the overhead of the access control is $O(N)$ (N denotes the number of the users); and the information exchange is $O(1)$. Therefore, the overall overhead of the proposed method is $O(N + 2)$.

From a functionality perspective, the content assignment part is responsible for acquiring the feedback of the received signal intensity of each user. Specifically, it first calculates the optional allocated content for the new user, minimizing the total multimedia distortion based on this feedback. To obtain an online solution, the dynamic programming algorithm introduced in [9] can be utilized directly in mmWave communication scenarios. All users are first assumed to have the same physical conditions (e.g., signal fading and multipath effect). Then all users are split into several sub-groups according to their physical conditions, and usually users with similar physical conditions form a sub-group. If the estimated multimedia distortion is smaller than the current one (i.e., with the group before splitting), the content assignment algorithm updates the set of groups. These steps are repeated until all the groups lead to minimizing the total multimedia distortion of all users. The advantage of this dynamic programming algorithm is the guaranteed optimality solution.

Multimedia Applications

As discussed previously, mmWave communication provides a broad platform for high-data-rate and short-distance multimedia applications. In this section we concentrate on the potential applications and their influences on our daily lives.

Potential Applications

Similar to UWB, mmWave is also suitable for high-data-rate and short-distance multimedia services, but it is subject to less intersystem interference. In general, previous research work found that mmWave can offer numerous applications in residential communities, universities, conference rooms, vehicles, and so on. In particular, it is appropriate for indoor communications such as video transmissions, device connections, and support of device-to-device communications. Multimedia applications can be divided into the following categories:

- Video/audio streaming
- File transfer
- Wireless gigabit ad hoc networks
- Wireless data service databases

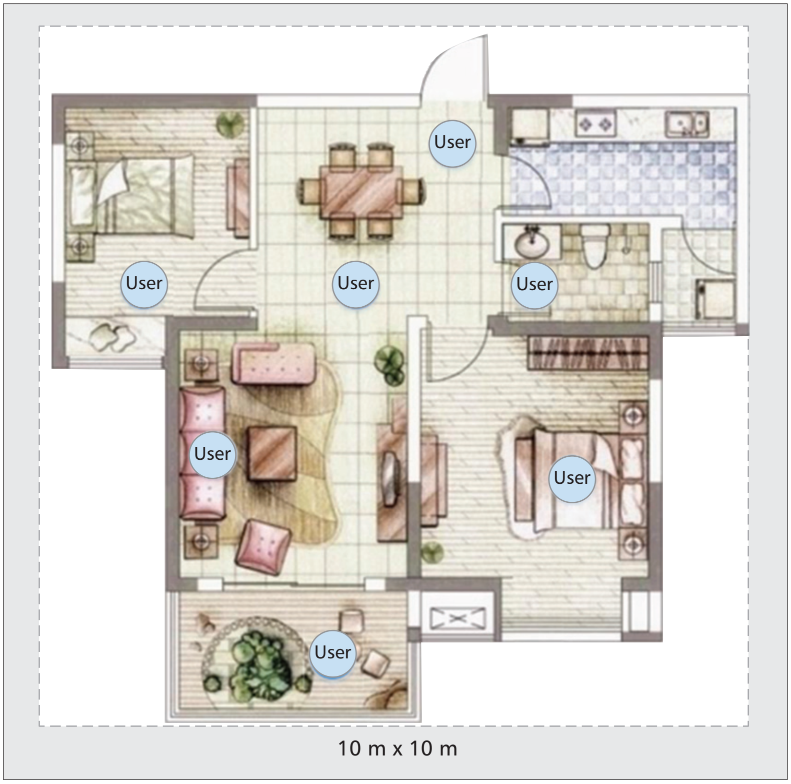


Figure 3. A practical mmWave environment.

Specifically, video streaming includes uncompressed/compressed versions for entertainment or data services. Usually, a video/audio stream is sent from a computer/laptop to a video/audio player. Typical transmission distance is less than 50 m, so it is possible to use it for high-bit-rate traffic. Moreover, multimedia streams can also broadcast from mobile devices arbitrarily placed in a room. In the home or office, any high-bit-rate file or data can be transmitted through any device in the network (printers, tablets, phones, computers, etc.), which can also be viewed as a wireless gigabit ad hoc network. In addition, multimedia streaming can also be used for a device-to-projector connection in a conference room where conference participants can freely share or show their files without switching the cable. For example, [1] develops a framework for performance characterization of short-range communications systems with the intention of investigating the feasibility of new multimedia wireless services at millimeter waves; [8] presents the results of the investigation of an indoor mmWave channel at 60 GHz in application to perspective WLAN systems; and [4] proposes an intra-car multimedia communications system using the 60 GHz band to achieve data rates of up to 1 Gb/s.

Service Mode

In the framework of mmWave communications, the above potential multimedia applications can be provided by two primary methods: peer-to-peer (P2P) streaming and adaptive HTTP streaming.

In P2P streaming, each user is both a content provider and a consumer. Scalable multimedia content enables each user to request only the available bit layers that can be supported by the transmission capacity. In contrast to conventional P2P multimedia service, there is a strict transmission delay constraint, which is usually defined as the waiting time between the service request and accomplishment for P2P streaming. To this end, P2P streaming systems typically employ a scalable time window for each part of the multimedia content that is relevant to the QoS level at the receiver. Note that the scalability of the window is dependent on the amount of transmission data and the constraint of the transmission delay. Within a time window, a media-aware scheduling algorithm, such as in [8], can be implemented at the receiver to minimize the content distortion or consumed energy.

In HTTP streaming, the transmitter should also provide information on multimedia content structure and representation. In particular, content representation represents the coding methods and parameters (frame rate, quantization step, etc.). For the layer-based coding method, this information also contains the structure of each bit layer. For the receiver, it deliberately chooses some of the information to decode the received multimedia content based on the given QoS requirement, computational ability, and transmission capacity. From the above description, although HTTP is suitable for unicast communication, it can also be applied to multicast streaming. In particular, for the multi-path scenario in mmWave communications, each path can be viewed as an independent unicast scenario. Although this method will inevitably lead to data redundancy, it will dramatically increase the reliability of the communication quality.

Illustrative Results

In this section we use numerical simulations to demonstrate the efficiency of the proposed multimedia scheduling scheme in a practical mmWave environment, as shown in Fig. 3. Two kinds of multimedia applications, audio and video, are employed. Specifically, the audio application is encoded with G.711 voice codec at 64 kb/s, while the video sequence "foreman" is encoded with the H.264 reference software encoder at 60 frames/s at a high-definition resolution (1280 × 720 pixels) [8]. We inject random background traffic at a rate between 30 and 50 percent of its available bit rate (ABR), which can be precisely detected in a practical scenario. The delay constraint is a random variable that-varies from 10 s to 300 s, and the feedback frequency of each user is 1 s.

The proposed QoS-aware multimedia scheduling scheme (called the QoS scheme) is benchmarked against the traditional distortion-driven scheduling proposed in [8] (called the Distortion method). In addition, in order to provide a clear picture of the performance, both the QoS and Distortion schemes are compared to a practical scenario where the two carrier frequencies, 60 and 70 GHz, are available, and the average transmission distance between each user varies from 1 to 10 m. In order to provide a fair comparison for both audio and video applica-

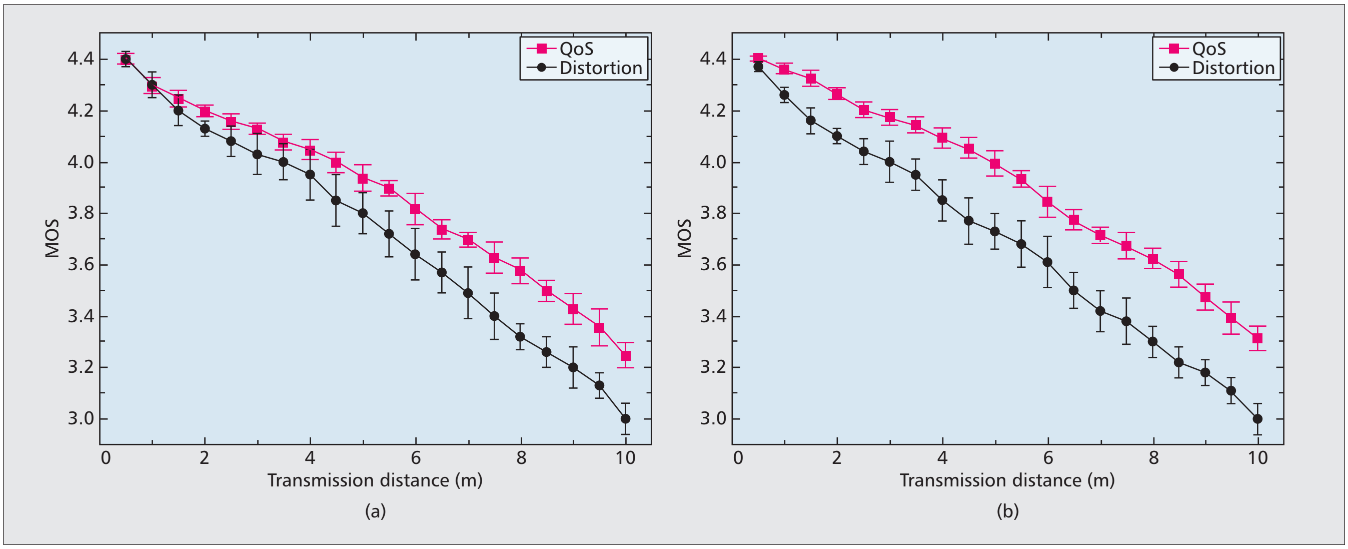


Figure 4. Performance comparison with different scenarios: a) 60 GHz; b) 70 GHz.

tions, we introduce the concept of mean opinion score (MOS), which reflects the degree of user satisfaction from a scale of 1 (unacceptable) to 4.5 (excellent) [9].

In Fig. 4 we show the average MOS value of the performance of the QoS and Distortion schemes with different simulation settings. From the given results, we can observe that:

- The QoS scheme outperforms the Distortion scheme, as expected, all the time. This is our main motivation for using user feedback.

2 The Distortion scheme is more sensitive to transmission distance than the QoS scheme. For example, the performance gap becomes larger when the transmission distance gets closer to 10 m. This is because the QoS scheme deploys the multimedia content estimation strategy to reduce the impact of the transmission distance.

3 The carrier frequency is significantly important for system performance. For example, at 60 GHz the average MOS value of the QoS scheme is 3.87, while it is 3.91 at 70 GHz.

As a result, the higher the carrier frequency the better the performance, which is consistent with the previous theoretical analysis.

Conclusion

This article aims to define and evaluate important metrics to characterize multimedia QoS for the user, and jointly takes into account several technical challenges, such as large-scale attenuation, atmospheric absorption, phase noise, and limited gain amplifiers. To this end, we designed a QoS-aware multimedia scheduling scheme to achieve a trade-off between performance and complexity, in which accurate propagation analysis is carried out and suitable countermeasure techniques are pointed out. Moreover, potential multimedia applications are analyzed and possible solutions provided. Illustrative results indicate that the proposed scheme can perform efficiently in a practical mmWave communication system.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61301163, 61322104, 61201165), and Jiangsu Provincial Natural Science Foundation of China (Grant BK20130067).

References

[1] O. Andrisano, V. Tralli, and R. Verdone, "Millimeter Waves for Short-Range Multimedia Communication Systems," *Proc. IEEE*, vol. 86, no. 7, 1998, pp. 1383–1401.

[2] R. C. Daniels and R. W. Heath, "60 GHz Wireless Communications: Emerging Requirements and Design Recommendations," *IEEE Vehic. Tech. Mag.*, vol. 2, no. 3, 2007, pp. 41–50.

[3] N. Guo *et al.*, "60-GHz Millimeter-Wave Radio: Principle, Technology, and New Results," *EURASIP J. Wireless Communications and Networking*, doi:10.1155/2007/68253, 2007.

[4] H. Sawada *et al.*, "A Sixty GHz Vehicle Area Network for Multimedia Communications," *IEEE JSAC*, vol. 27, no. 8, 2009, pp. 1500–06.

[5] T. S. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," *IEEE Access*, vol. 1, no. 1, 2013, pp. 335–49.

[6] L. Zhou *et al.*, "Fairness Resource Allocation in Blind Wireless Multimedia Communications," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 946–56, June 2013.

[7] J. Yang *et al.*, "A Framework for Classifier Adaptation for Large-Scale Multimedia Data," *Proc. IEEE*, vol. 100, no. 9, 2012, pp. 2639–57.

[8] A. Maltsev *et al.*, "Characteristics of Indoor Millimeter-Wave Channel at 60 GHz in Application to Perspective WLAN System," *Proc. 4th Euro. Conf. Antennas and Propagation*, 2010.

[9] L. Zhou *et al.*, "Distributed Wireless Video Scheduling with Delayed Control Information," *IEEE Trans. Circuits and Sys. for Video Tech.*, vol. 24, no. 5, 2014, pp. 889–901.

Biographies

Dan Wu (wujing1958725@126.com) received her B.S., M.S., and Ph.D. degrees at the Institute of Communications Engineering, PLA University of Science and Technology, Nanjing, China in 2006, 2009, and 2012, respectively. She is now a postdoctoral researcher at the Institute of Communications Engineering, PLA University of Science and Technology, Nanjing, China. Her research interests are mainly in resource allocation and management, game theory, cooperative communications, and wireless sensor networks.

Jinlong Wang (wjl543@sina.com) received the B.S. degree in wireless communications and the M.S. and Ph.D. degrees in communications and electronic systems from the PLA

University of Science and Technology, Nanjing, China, in 1983, 1986, and 1992, respectively. He is currently a professor with the Institute of Communications Engineering, PLA University of Science and Technology. He has widely published in signal processing for communications, information theory, and wireless networks. His research interests include wireless communication, cognitive radio, soft-defined radio, and ultrawide bandwidth systems. He is also a co-chairman of the Institute of Electrical and Electronics Engineers Nanjing Section.

YUMING CAI (caiym@vip.sina.com) received his B.S. degree in physics from Xiamen University, Xiamen, China in 1982, the M.S. degree in micro-electronics engineering and the Ph.D. degree in communications and information systems both from Southeast University, Nanjing, China in 1988 and 1996, respectively. His current research interests include MIMO systems, OFDM systems, signal processing in communications, cooperative communications, and wireless sensor networks.

MOHSEN GUIZANI (mguizani@ieee.org) is currently a professor and the associate vice president for Graduate Studies at Qatar University, Qatar. He was the chair of the CS Department at Western Michigan University from 2002 to 2006, and chair of the CS Department at the University of West Florida from 1999 to 2002. He also served in academic positions at the University of Missouri-Kansas City, University of Colorado-Boulder, Syracuse University, and Kuwait University. He received his B.S. (with distinction) and M.S. degrees in electrical engineering, and the M.S. and Ph.D. degrees in computer engineering in 1984, 1986, 1987, and 1990, respectively, from Syracuse University, Syracuse, New York.

His research interests include security, smart grid, wireless communications and mobile computing, and optical networking. He currently serves on the editorial boards of six technical journals, and is the founder and EIC of the journal *Wireless Communications and Mobile Computing*, published by John Wiley (http://www.interscience.wiley. com/jpages/1530-8669/). He is also the founder and Steering Committee Chair of the Annual International Conference of Wireless Communications and Mobile Computing (IWCMC). He is the author of seven books and more than 400 publications in refereed journals and conferences. He has guest edited a number of special issues in IEEE journals and magazines. He has also served as member, chair, and general chair of a number of conferences. Dr. Guizani served as the chair of the IEEE Communications Society Wireless Technical Committee (WTC) and chair of the TAOS Technical Committee. He was an IEEE Computer Society Distinguished Lecturer from 2003 to 2005. Dr. Guizani is an IEEE Fellow and a Senior Member of ACM.

Call for Papers
IEEE Communications Magazine
Green Communications and Computing Networks Series

Background

Green Communications and Computing Networks is issued semi-annually as a recurring Series in *IEEE Communications Magazine*. The objective of this Series is to provide a premier forum across academia and industry to address all important issues relevant to green communications, computing, and systems. The Series will explore specific green themes in depth, highlighting recent research achievements in the field. Contributions provide insight into relevant theoretical and practical issues from different perspectives, address the environmental impact of the development of information and communication technologies (ICT) industries, discuss the importance and benefits of achieving green ICT, and introduce the efforts and challenges in green ICT. This Series welcomes submissions on various cross-disciplinary topics relevant to green ICT. Both original research and review papers are encouraged. Possible topics in this Series include, but are not limited to:

- Green concepts, principles, mechanisms, design, algorithms, analyses, and research challenges
- Green characterization, metrics, performance, measurement, profiling, testbeds, and results
- Context-based green awareness
- Energy efficiency
- Resource efficiency
- Green wireless and/or wireline communications
- Use of cognitive principles to achieve green objectives
- Sustainability, and environmental protections by and for ICT
- ICT for green objectives
- Non-energy-relevant green issues and/or approaches
- Power-efficient cooling and air conditioning
- Green software, hardware, device, and equipment
- Environmental monitoring
- Electromagnetic pollution mitigation
- Green data storage, data centers, contention distribution networks, and cloud computing
- Energy harvesting, storage, transfer, and recycling
- Relevant standardizations, policies, and regulations
- Green smart grids
- Green security strategies and designs
- Green engineering, agenda, supply chains, logistics, audit, and industrial processes
- Green building, factory, office, and campus designs
- Application layer issues
- Green scheduling and/or resource allocation
- Green services and operations
- Approaches and issues of social networks used to achieve green behaviors and objectives
- Economic and business impact and issues of green computing, communications, and systems
- Cost, OPEX, and CAPEX for green computing, communications, and systems
- Roadmap for sustainable ICT
- Interdisciplinary green technologies and issues
- Recycling and reuse
- Prospect and impact on carbon emissions and climate policy
- Social awareness of the importance of sustainable and green communications and computing

Submission Guidelines

Prospective authors are strongly encouraged to contact the Series Editor with a brief abstract of the article to be submitted before writing and submitting an article in order to ensure that the article will be appropriate for the Series. All manuscripts should conform to the standard format as indicated in the submission guidelines at

http://www.comsoc.org/commag/paper-submission-guidelines

Manuscripts must be submitted through the magazine's submissions web site at

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

You will need to register and then proceed to the Author Center. On the manuscript details page, please select "Green Communications and Computing Networks Series" from the dropdown menu.

Publication Schedule

Inaugural Issue: November 2014
Scheduled Publication Dates: Twice per year, May and November

Series Editors

Jinsong Wu, Alcatel-Lucent, China, wujs@ieee.org
John Thompson, University of Edinburgh, UK, john.thompson@ed.ac.uk
Honggang Zhang, UEB/Supelec, France; Zhejiang Univ., China, honggangzhang@zju.edu.cn
Daniel C. Kilper, University of Arizona, USA, dkilper@optics.arizona.edu

Multimedia Resource Allocation in mmWave 5G Networks

Sandra Scott-Hayward and Emiliano Garcia-Palacios

The authors are with Queen's University Belfast.

ABSTRACT

The 5G network infrastructure is driven by the evolution of today's most demanding applications. Already, multimedia applications such as on-demand HD video and IPTV require gigabit-per-second throughput and low delay, while future technologies include ultra HDTV and machine-to-machine communication. Mm-Wave technologies such as IEEE 802.15.3c and IEEE 802.11ad are ideal candidates to deliver high throughput to multiple users demanding differentiated QoS. Optimization is often used as a methodology to meet throughput and delay constraints. However, traditional optimization techniques are not suited to a mixed set of multimedia applications. Particle swarm optimization (PSO) is shown as a promising technique in this context. Channel-time allocation PSO (CTA-PSO) is successfully shown here to allocate resource even in scenarios where blockage of the 60 GHz signal poses significant challenges.

INTRODUCTION

The volume of mobile traffic is exploding, driven by a proliferation of connected devices. The high bandwidth required by multimedia applications is severely stretching the available wireless spectrum. Consumers influenced by fixed access broadband services expect almost instant file uploads/downloads and transfers between terminals (laptops, tablets, and smart TVs), IPTV, and HD video. The challenge for service providers is therefore to extend the capability of wireless access to maintain performance characteristics in the transition to predominantly wireless service provision. The entertainment industry will be just the first adopter to drive and inspire other futuristic scenarios in connected health, social and community interaction, and education.

The 60 GHz band, with up to four 2.16 GHz channels, supports high-data-rate short-range line-of-sight (LOS) directional transmissions. Millimeter-wave (mmWave) communication technologies such as IEEE 802.15.3c-2009 [1] and IEEE 802.11ad [2] are uniquely positioned due to their ability to deliver the gigabit-per-second throughput envisaged for the fifth generation (5G).

Despite the adverse propagation characteristics of mmWave technologies (a result of high propagation loss due to oxygen absorption and atmospheric attenuation), the fact that waves are confined within walls makes it ideal for personal communications: promoting security and privacy. This also promotes more efficient frequency reuse, therefore creating very high bandwidth hubs that are ideal for the delivery of gigabit-per-second applications. A major question, however, is how to combat the impact of blockage of the LOS link by human shadowing or obstacles, which severely impacts the transmission.

In order to support the high volume of high-quality services, optimized resource allocation is required. Optimizing resource allocation in a multi-user multimedia gigabit scenario poses interesting challenges. Most traditional optimization techniques have been used to solve convex optimization problems. However, real-time multimedia applications introduce non-convex utility functions in which the quality perceived by the user does not increase gracefully with an increase in throughput. A re-think of resource allocation optimization techniques is necessary in this context. Relatively new approaches, such as particle swarm optimization (PSO), are ideal candidates to solve non-convex problems. Our proposal compares PSO to other techniques and presents channel-time allocation PSO (CTA-PSO) as a solution to optimize channel time allocation (resource) even when blockage occurs in the mmWave network.

ENABLING MMWAVE TECHNOLOGIES

IEEE 802.15.3c [1] and IEEE 802.11ad [2] provide mechanisms at the medium access control (MAC) layer to allocate resource (in the form of channel access time) to multiple users. A period of contention allows for different terminals to request access time while allocated channel time slots provide dedicated data transmission time. The advantage of scheduled access is to provide a dedicated time slot for communication, thus supporting guaranteed quality of service (QoS) for an application. This guarantee is not possible when all network devices compete for bandwidth using random access techniques. The allocation process is described here and illustrated in Fig. 1.

0163-6804/15/$25.00 © 2015 IEEE

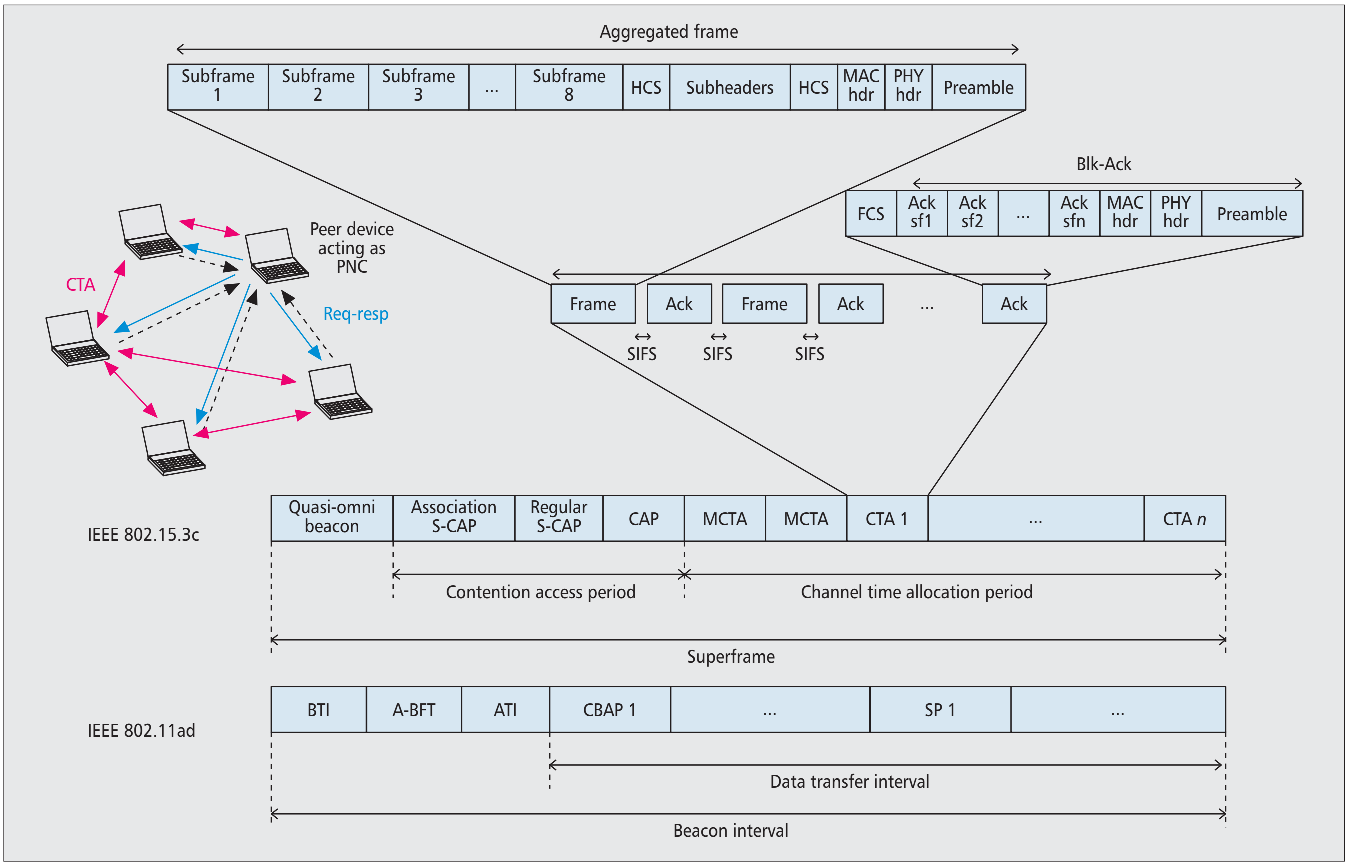


Figure 1. MAC layer structure in IEEE 802.11ad and IEEE 802.15.3c

Allocating Resource in IEEE 802.15.3c

An 802.15.3c network, or piconet, is a wireless ad hoc data communications system in which a number of independent devices communicate with each other in a peer-to-peer fashion [1]. In the 802.15.3c protocol, medium access is controlled by a piconet controller (PNC). This role is usually held by the first member of the network. The superframe is initiated by a beacon from the PNC. In order to achieve the required signal range at 60 GHz, a directional beacon is employed, but in order to reach all potential neighbors, the beacon must be transmitted in all directions. A quasi-omni beacon is therefore used. The area around the PNC is divided into sectors, and the PNC transmits a directional beam beacon sector by sector until all sectors have been covered.

A contention access period (CAP) follows the beacon phase. The S-CAP is sectorized, so in the association S-CAP, all devices in a sector have the opportunity to contend to send association request commands and for the PNC to send immediate acknowledgments. The regular S-CAP and CAP are then available for command and data exchanges.

The CAP is a carrier sense multiple access with collision avoidance (CSMA/CA) phase. In CSMA/CA, a device first senses the channel prior to transmission. If the channel is free, the device transmits its frame. Otherwise, if the channel is sensed busy, the device generates a random backoff counter and defers its transmission until the backoff counter expires.

In 802.15.3c, devices compete during the CAP to transmit a request for dedicated channel time during which the application data can be transmitted. The device time allocation is called a channel time allocation (CTA) and occurs during the channel time allocation period (CTAP), which follows the CAP. The CTAP operates with time-division multiple access (TDMA). In this phase, a time slot/CTA is assigned to any devices that have requested an access period, provided the time is available.

In order to improve the MAC efficiency, frame aggregation and block acknowledgment are supported by the standard. These techniques are also illustrated in Fig. 1. Frame aggregation involves mapping MAC service data units (MSDUs) into multiple subframe payloads. This increases the data to header ratio in the frame thus increasing throughput. A block acknowledgment (Blk-Ack) is sent for the aggregated frame. The benefit of these mechanisms is improved efficiency by reducing frame/Ack overhead.

Allocating Resource in IEEE 802.11ad

Accommodating the constraints of the mmWave frequency band and supporting the requirements of high-data-rate applications, IEEE 802.11ad supplements and extends earlier versions of IEEE 802.11 MAC. It operates in a similar way to 802.15.3c with periods of contention-based and scheduled access. However, the allocation of

Optimization involves finding the best solution to a problem. Mathematically, this involves maximization/minimization of an objective/utility function, which may be unconstrained or subject to certain constraints on the variable(s) of the function.

these transmission periods differs from the 802.15.3c approach, as illustrated in Fig. 1.

Channel access time in IEEE 802.11ad is divided into beacon intervals (BIs). Each BI consists of the beacon transmission interval (BTI), association beamforming training (A-BFT), announcement transmission interval (ATI), and data transfer interval (DTI). During the BTI, the control/access point transmits directional beacons to each sector of the sectorized network. Request-response-based transmission allocations of scheduled service periods (SPs) and contention-based access periods (CBAPs) take place in the ATI. Finally, the DTI consists of a series of SPs and CBAPs during which application data transmissions occur.

The importance of the compatibility of IEEE 802.11ad with legacy 802.11 standards is clear. A transition will be possible between the lower frequency band (2.4/5 GHz) for long-range communication and the higher frequency band (60 GHz) for short-range communication [3]. It is this heterogeneous network style that will underpin the 5G infrastructure.

While both protocols described here combine CSMA/CA for contention-based medium access and TDMA for scheduled service, they differ in the order of allocation. In IEEE 802.15.3c, the CTA requests are sent during an initial contention-based period (CAP), and the CTAP consumes the majority of the superframe with individual time slots (CTAs) allocated for application data transmission. In contrast, contention-based access periods (CBAPs) and scheduled service periods (SPs) are alternated in IEEE 802.11ad to support compatibility with legacy 802.11 standards. The objective of both methods is to maximize network throughput. Optimizing resource allocation is key to maximizing network throughput.

One approach to resource allocation for multiple applications is the assignment of packets to queues, where each queue has a priority level for accessing the wireless medium. For example, the packets of a real-time low-latency application would receive the highest priority in accessing the available bandwidth. However, although this achieves prioritization between application types, it does not consider the specific (and variable) requirements of individual devices within an application type and how throughput/QoS might be optimized by appropriate scheduling of these transmissions. It is our belief that both the application type and the specific individual application requirements must be considered for optimal resource allocation.

In our consideration of the channel time allocation optimization problem, we exploit the MAC layer structure of IEEE 802.15.3c and IEEE 802.11ad. The contention periods are used for the exchange of time slot request-response messages. The contention-free/scheduled service periods are dedicated to application data transmission with defined QoS.

Optimization Techniques

Optimization involves finding the best solution to a problem. Mathematically, this involves maximization/minimization of an objective/utility function, which may be unconstrained or subject to certain constraints on the variable(s) of the function.

The nature of the utility function for multimedia applications determines the type of optimization problem [4]. Classically, resource management solutions handle video as an isolated application in the network, thus solving a convex optimization problem. For example, the higher the data rate at which the video can be transmitted, the better the quality perceived by the user and hence the higher the utility for the user.

In a practical network, however, mixed multimedia applications (e.g., IPTV, VoIP) must be served along with video. The utility functions to describe such applications are non-convex. The non-convexity arises from the inelasticity of real-time applications, meaning that the application does not improve or degrade gracefully in response to an increase or a reduction in allocated transmission rate [5]. Rather, a reduction in data rate below a certain threshold results in a significant drop in QoS.

The challenge for 5G is therefore a combination of resolving the non-convex optimization problem (resulting from multiple applications) within the constraints of the mmWave network. This includes the practical consideration of the execution time of the algorithm. Framing resource allocation as a network utility maximization (NUM) problem, the implementation of four popular optimization techniques are compared here for their solution potential. Techniques suitable for convex optimization problems only are included in order to illustrate the difference in attributes of each technique, specifically with respect to speed of execution of the algorithm.

The problem is defined in Eq. 1.

$$\begin{gathered}\text{Maximize}\sum_{i=1}^{N} U_i(CTA_i)\\ \text{Subject to}\sum_{i=1}^{N} CTA_i \le CTAP\\ CL_i \le CTA_i \le CH_i \quad \forall\ i = 1, 2, \ldots, N\end{gathered} \tag{1}$$

NUM is the problem of maximizing the total utility, U_i, of the network over the channel time allocations, **CTA**, subject to the constraint that the sum of all CTAs should not exceed the network capacity (CTAP). In addition, upper and lower CTA limits, $[CL_i, CH_i]$, are set based on the desired quality constraint of the individual application. The result of the NUM will be a time slot (CTA) allocation for each application, which takes into account both the physical transmission rate on each wireless link based on the channel condition and the immediate quality requirements of the application.

The optimization techniques are:

- Lagrangian Dual Decomposition Subgradient Algorithm
- Rate allocation game (Nash equilibrium via pricing mechanism)
- Nash bargaining solution
- Particle swarm optimization

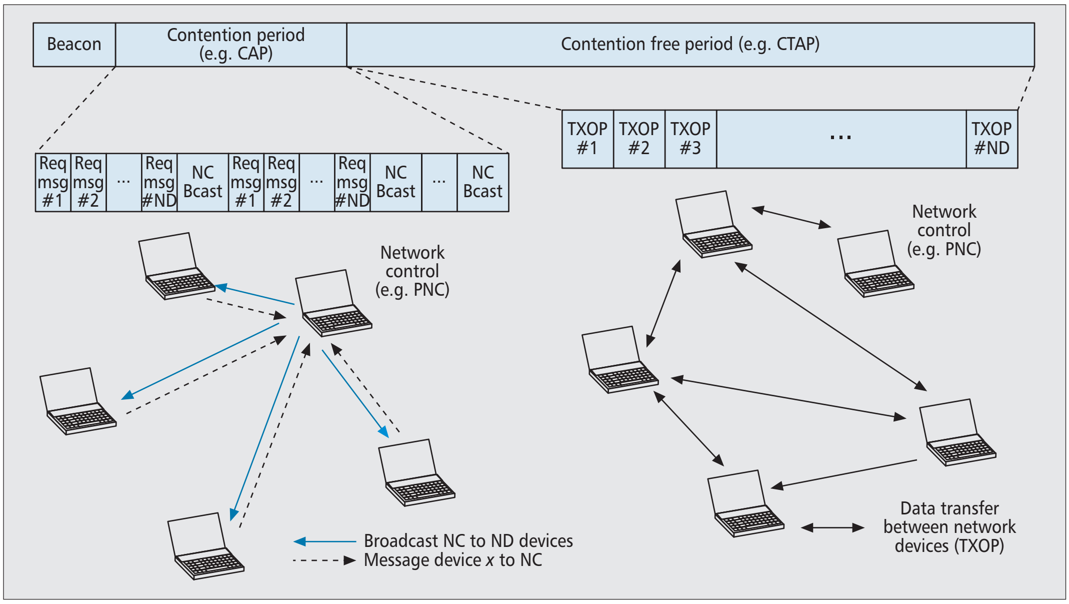

Beacon
Contention period (e.g. CAP)
Contention free period (e.g. CTAP)
Req msg #1
Req msg #2
...
Req msg #ND
NC Bcast
Req msg #1
Req msg #2
...
Req msg #ND
NC Bcast
...
NC Bcast
TXOP #1
TXOP #2
TXOP #3
...
TXOP #ND
Network control (e.g. PNC)
Network control (e.g. PNC)
Broadcast NC to ND devices
Message device x to NC
Data transfer between network devices (TXOP)

Figure 2. Implementation of the rate allocation game in the hybrid MAC framework: the request/response exchange for transmission opportunities (TXOPs) takes place in the CAP with the TXOPs provided in the CTAP.

Classically, resource management solutions handle video as an isolated application in the network, thus solving a convex optimization problem. For example, the higher the data rate at which the video can be transmitted, the better the quality perceived by the user and hence the higher the utility for the user.

Lagrangian Dual Decomposition Subgradient Algorithm

The first optimization method considered is the Lagrangian dual decomposition subgradient algorithm. This evolved from the gradient search methods used in convex optimization. With gradient search, the search direction is defined by the gradient of the function to be optimized at the current point, the gradient being the first derivative of the function. For a convex programming problem, the gradient projection controlled by the Lagrange multiplier leads to convergence to the optimal solution. However, gradient search methods are inapplicable to non-differentiable convex optimization problems.

A subgradient approach to the Lagrangian algorithm presented in [6] enables solution of non-differentiable convex problems. An iterative algorithm is generated, replacing the gradient-based method. The Lagrange multiplier (or price per unit rate in the resource allocation problem) is updated at each iteration. The implementation of the algorithm requires message passing in order to communicate the Lagrange multiplier from each network device to a central calculation point at each iteration. The disadvantage of message passing is the overhead introduced to the network, which reduces application data transmission time.

The Rate Allocation Game

Game theory is a set of mathematical tools used to analyze interactive decision processes [7]. The rate allocation game is non-cooperative in the sense that each device acts as a selfish player in the game. This enables distributed implementation. However, the use of a pricing mechanism handles the conflicting objectives of the wireless devices in the network.

The game is described by a number of devices (players), a vector of strategies/actions, and a vector of payoffs based on the strategies chosen by the devices in the network.

The players are rational. This means that they take account of the consequences of their choices and of other players' choices in order to selfishly maximize their utility. As described in [8], a rational player will use only those strategies that are best responses to some beliefs she might have about the strategies of her opponents. The requests of other devices are taken into account in a cost term describing the price of requesting additional transmission rate. If there are few devices competing for the available resource, the price is low. However, if there is high competition, the price for requesting more resource is higher. If price increases, satisfaction decreases, so an optimum is reached where no user wishes to deviate because their satisfaction will be reduced if the price goes up, which is the consequence of excess demand. The payoff reflects the overall loss/gain the player incurs based on its selected strategy. The resulting resource allocation is a Nash equilibrium [8]. An illustration of how the game would be implemented in an mmWave network is provided in Fig. 2.

Nash Bargaining Solution

The Nash bargaining solution (NBS) is another game-theoretic method. In this case, players cooperate to reach a fair allocation of resources. Each player has a minimum resource acceptable to it, known as the disagreement point, d. NBS allocates resources optimally by maximizing the Nash product, which is the product of utilities. In the channel time allocation problem, $\mathbf{d}$ corresponds to the set of lower CTA limits, **CL**.

NBS is an axiomatic bargaining solution, which means that it does not require iterative bargaining among users, consequently avoiding message passing. However, neither NBS nor the rate allocation game is applicable to non-convex functions.

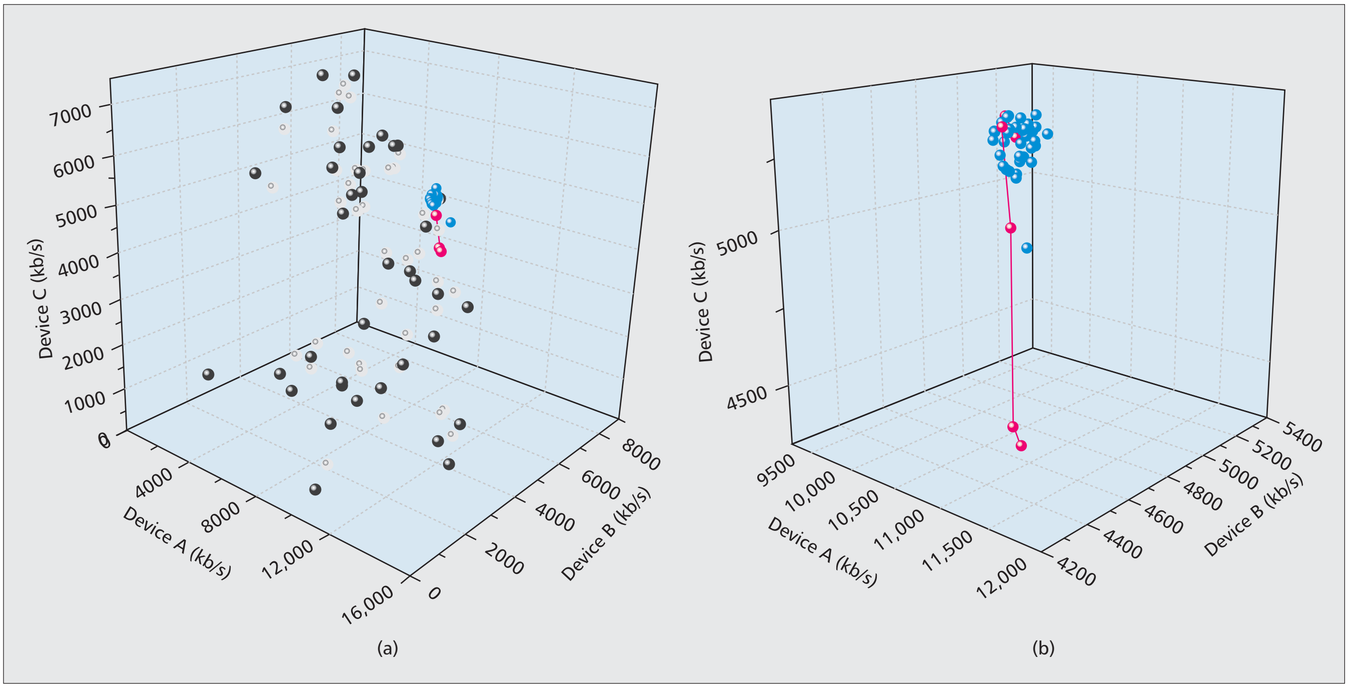


Figure 3. PSO convergence illustration in a 3D search space: a) first (black) and fifth (grey) iterations; b) final (cyan) iteration with indication of progression toward *Gbest* (red).

Particle Swarm Optimization

Direct search methods or evolutionary techniques can be employed to solve non-convex problems. A range of evolutionary techniques have been developed in the past ~40 years, the first of which was the Genetic Algorithm (GA) developed in 1975 by John Holland and his students.

In 1995, Kennedy and Eberhart introduced their PSO algorithm for solving global optimization problems [9]. Of the evolutionary algorithms, PSO has the appeal of simplicity and evidence of good performance in a variety of application domains. PSO is based on social behavior with multiple potential solutions of a problem generated at initialization. The solution set is called a swarm, and each solution is a particle. The particles move in the problem search space seeking the optimal solution. In a similar manner to the social and cooperative behavior of species like birds and fish, they exchange their knowledge of the search space to find the best solution by self-learning and collaboration. At each iteration, each particle adjusts its position according to its own experience and the experience of its neighbors. It is a centralized algorithm suitable for solution of non-convex problems.

A 3D PSO for a three-device resource allocation problem is presented in Fig. 3. The convergence of the particles toward the global best position, *Gbest*, is illustrated. In Fig. 3a, the distribution of the swarm of 40 particles across the search space in the early PSO iterations is shown. In Fig. 3b, a zoom-in on this search space highlights the connecting line indicating the progression of the global best position over a number of iterations and clustering of the particles around the final *Gbest*, which is the resource allocation solution for the three-device network.

Evaluation of Optimization Techniques

The attributes of each optimization technique are outlined in Table 1. The execution time is ordered in terms of increasing requirement from 1 to 4 based on an example execution of 8 devices each running a different video sequence.

The rate allocation game is limited by both its high message-passing overhead and its constraint to convex problems. Although the fastest in terms of execution time, the NBS is also limited to convex optimization problems. As indicated in Table 1, a Lagrangian relaxation approach could be explored for the non-convex resource allocation problem. However, the overhead introduced by message passing proves a limitation of this approach.

In contrast, if the time required to reach convergence could be reduced, the benefit of centralized implementation of PSO along with its capacity to resolve non-convex optimization problems presents a potential method of solving the multimedia resource allocation problem.

Channel Time Allocation Particle Swarm Optimization

The results of Table 1 indicate that PSO has potential to resolve the wireless multimedia network resource allocation problem.

PSO has the appeal of simplicity and evidence of good performance in a variety of application domains. It has been demonstrated to be more computationally efficient than GA on a series of test problems [10].

With the focus of achieving a near global optimal solution in a time suitable for implementation in a real network, CTA-PSO has been developed [4]. CTA-PSO overcomes premature convergence by controlling exploration and

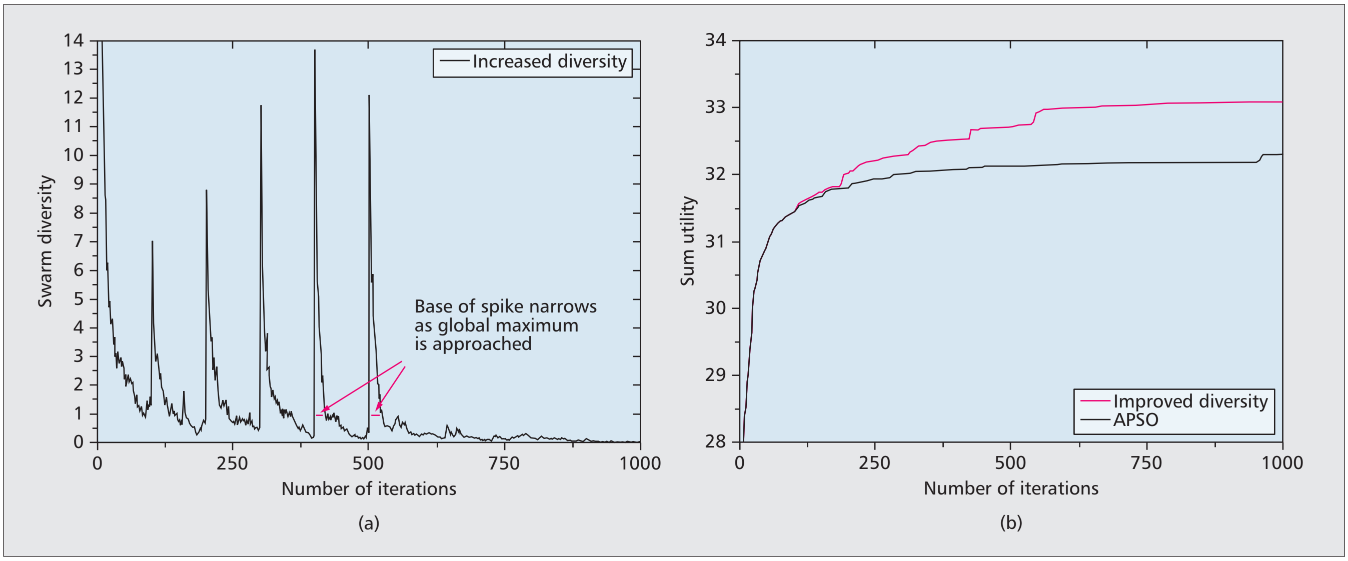


Figure 4. CTA-PSO 40-device example: a) swarm diversity; b) convergence behavior.

exploitation in the search space. Exploration refers to the ability of the swarm to explore different regions of the search space in order to locate the global optimum. Exploitation refers to the ability of the particles to concentrate the search around a promising area of the search space in order to refine a potential solution.

The parameters in the PSO equation (Eq. 2) contribute to this.

$$V_i^{t+1} = \omega V_i^t + c_1 r_1 (Pbest_i^t - CTA_i^t) + c_2 r_2 (Gbest_i^t - CTA_i^t)$$
$$CTA_i^{t+1} = CT\,A_i^t + V_i^{t+1} \qquad (2)$$

At each iteration, each particle's velocity, V_i, and channel time allocation, CTA_i, is updated. The dimension of CTA_i is equal to the number of devices to be allocated resource in the network.

In Eq. 2, t is the iteration number, r_1/r_2 are uniform random numbers, and c_1/c_2 are learning rates/acceleration coefficients representing the weight of memory of a particle's best position, *Pbest*, toward the memory of the swarm best position, *Gbest*. ω is the inertia weight, controlling the contribution of the previous velocity to the velocity update.

A particle keeps track of its coordinates in the search space and aims to reach *Gbest*. The best solution is determined by the value of the fitness function, which in the resource allocation problem of Eq. 1 is the utility function to be maximized.

The PSO fitness function, F, for the CTA problem is described in Eq. 3.

$$F = \begin{cases} \sum_{i=1}^{N} U_i(CTA_i) & \text{if } \sum_{i=1}^{N} CTA_i \le CTAP, \\ \sum_{i=1}^{N} U_i(CTA_i) & \\ \quad + \gamma\left(CTAP - \sum_{i=1}^{N} CTA_i\right) & \textit{otherwise}, \end{cases} \qquad (3)$$

where the penalty value γ > 0. The penalty value accommodates the practical constraint that the sum time allocated must not exceed the available resource (i.e., the *CTAP*).

CTA-PSO monitors similarity in the swarm and achieves the global solution fast by introducing particles to increase diversity and reducing computation within the PSO. Increasing the diversity in the swarm avoids early convergence and stagnation of the swarm at a local maximum. Based on monitoring the diversity of the swarm at intervals, a percentage of particles with the worst fitness are removed and replaced with new particles. This has the effect of injecting new energy into the PSO to break out of the local maximum and search for the global maximum.

Furthermore, in the context of dynamic implementation of CTA-PSO in a wireless network, a learning element is introduced based on neighboring group of pictures (GoP) similarity.

Method	Execution time ranking	Convex functions	Non-convex functions	Centralized	Message passing required
Lagrangian dual decomposition	4	✓	✓	X	✓
Rate allocation game	2	✓	X	X	✓
Nash bargaining solution	1	✓	X	✓	X
Particle swarm optimization	3	✓	✓	✓	X

Table 1. Comparison of optimization methods for wireless multimedia resource allocation.

This feature takes account of the particular application type (e.g., video) and the observation that the GoP size remains approximately the same until a change in video scene takes place. As a result, rather than randomizing the swarm particles at each execution of the algorithm, knowledge of the previous best particle, *Pbest*, positions is used. With this *Pbest* learning approach, CTA-PSO execution time can be further reduced.

The variation in swarm diversity is shown in Fig. 4a. The spikes in the graph illustrate the implementation of the diversity function. The higher diversity values represent greater exploration by the swarm. The base of the spike narrows (spikes 4 and 5 in Fig. 4a) indicating increased localization of the swarm exploration and can be considered as confirmation that the global optimum has been found.

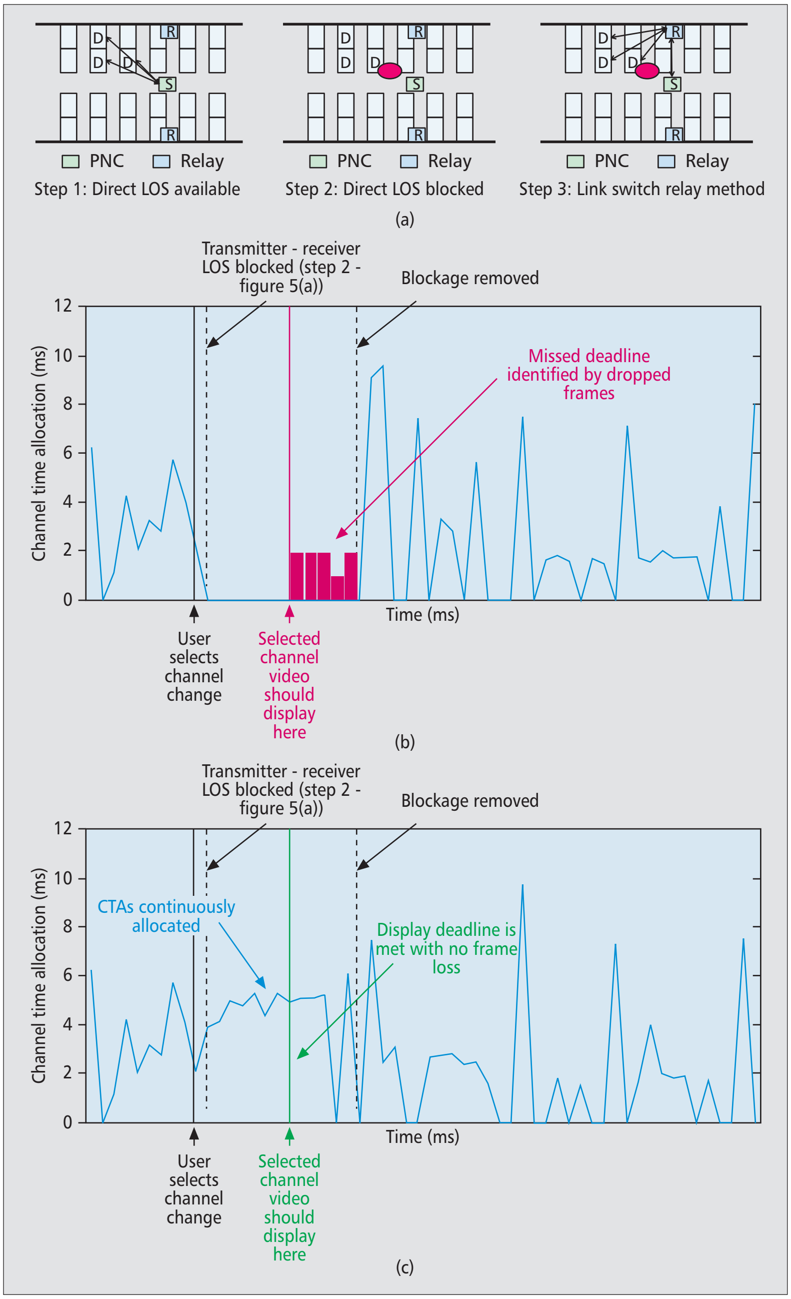


Figure 5. Impact of 1 s blockage on CC IPTV: a) blockage event; b) PSO resource allocation without link switch; c) PSO resource allocation with link switch during the blockage phase.

The corresponding utility curve is shown in Fig. 4b, illustrating the improvement in sum utility over an alternative recognized PSO algorithm, APSO [11].

Solving the Challenge of 60 GHz Resource Allocation Using CTA-PSO

We present an example where CTA-PSO will solve the resource allocation problem in the presence of blockage. The environment is an in-vehicle entertainment system, for example, within a train carriage where the blockage could be due to a person standing/walking in the aisle.

A real-time application such as live IPTV where a user suddenly changes to a new TV channel (IPTV channel change or CC IPTV) poses significant challenges when meeting delay deadlines imposed by this application (i.e., selected channel display deadline). The problem is aggravated by the fact that there are no IPTV frames previously buffered, and the direct LOS is blocked (Fig. 5a, step 2). The blockage is solved by means of the link switch relay method (IEEE 802.11ad) as shown in Fig. 5a, step 3. An IPTV utility function is designed to reflect the delay requirement to be met and minimize the number of IPTV frames being lost [12].

The advantages of "link switch" are clearly visible in comparing Figs. 5b and 5c. The user requests a channel change and shortly afterward a blockage occurs for a duration of 1 s (approximately 15 MAC super-frames). The display deadline computed to meet QoS demands is 720 ms (11 superframes) indicated in Figs. 5b and 5c by the "channel should display here" legend. This value is based on the difference between the maximum QoE-linked channel change delay (2 s) and the response time of the channel change request, which includes CTA-PSO and application-related timing values [12]. Without link switch there is no CTA time allocated during the blockage phase. As a result, the deadline is missed and frames are dropped. Frames are only transmitted again once the blockage is removed and CTA time is increased by PSO, as shown in Fig. 5b.

Using the link switch relay method solves the problem as shown in Fig. 5c. CTA-PSO adjusts immediately to the relay introduced in the communication path. As a result, CTA time is continuously allocated despite the direct LOS blockage, so the display deadline is met with zero frame loss.

Referring back to the methods introduced earlier, neither NBS nor RAG would be able to solve this non-convex optimization problem. For this practical situation, the increased implementation time of the Lagrangian optimization technique would make it challenging to allocate the appropriate resource to overcome the blockage.

This case study shows a typical and realistic scenario for mmWave applications where not only does access time have to be optimized, but signal blockage also has to be overcome. The combination of CTA-PSO with the delay-sensitive utility function and link switch provides a solution to this challenge.

CONCLUSION

Current multimedia applications (HD video, IPTV) and future technologies for smart homes, smart vehicles, and smart cities place high demands on wireless technologies. Smart resource allocation is required to achieve and maintain the QoS expected by the user.

To date, optimization techniques have focused on single application solutions. In order to meet the future multi-user multimedia heterogeneous network demands of 5G, alternative resource allocation optimization techniques must be explored.

In this article, we propose a reduced execution-time solution such as CTA-PSO, demonstrating its suitability for implementation in a challenging mixed multimedia wireless environment. In order to meet the evolving requirements of new applications, and converged and high-capacity networks such as 5G, alternative resource allocation techniques such as CTA-PSO must be further explored.

In order to meet the evolving requirements of new applications, and converged and high-capacity networks such as 5G, alternative resource allocation techniques such as CTA-PSO must be further explored.

REFERENCES

[1] IEEE 802.15.3c, "IEEE Standard for Information Technology — Telecommunications and Information Exchange between Systems — Local and Metropolitan Area Networks — Specific Requirements. Part 15.3: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Wireless Personal Area Networks (WPANs) Amendment 2: Millimeter-Wave-Based Alternative Physical Layer Extension," IEEE Std 802.15.3c-2009 (Amendment to IEEE Std 02.15.3-2003), 2009, pp. c1–187.

[2] IEEE802.11ad-2012, "IEEE Standard for Information Technology — Telecommunications and Information Exchange between Systems — Local and Metropolitan Area Networks — Specific Requirements. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60GHz Band," IEEE Std 802.11ad-2012 (Amendment to IEEE Std 802.11-2012), 2012, pp. 1–628.

[3] P. Smulders, "Exploiting the 60 GHz Band for Local Wireless Multimedia Access: Prospects and Future Directions," *IEEE Commun. Mag.*, vol. 40, no. 1, 2002, p. 140.

[4] S. Scott-Hayward and E. Garcia-Palacios, "Channel Time Allocation PSO for Gigabit Multimedia Wireless Networks," *IEEE Trans. Multimedia*, vol. 16, no. 3, 2014, pp. 828–36.

[5] S. Shenker, "Fundamental Design Issues for the Future Internet," *IEEE JSAC*, vol. 13, no. 7, 1995, pp. 1176–88.

[6] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, "Nonconvexity Issues for Internet Rate Control with Multiclass Services: Stability and Optimality," *Proc. 23rd Annual IEEE INFOCOM*, vol. 1, 2004, pp. 1–12.

[7] J. Neel, J. H. Reed, and R. P. Gilles, "Game Models for Cognitive Radio Algorithm Analysis," *Proc. SDR Forum Tech. Conf.*, Nov. 2004.

[8] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, 1991.

[9] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," *Proc. IEEE Int'l. Conf. Neural Net.*, vol. 4, 1995, pp. 1942–48.

[10] R. Hassan *et al.*, "A Comparison of Particle Swarm Optimization and the Genetic Algorithm," *Proc. 1st AIAA Multidisciplinary Design Optimization Specialist Conf.*, 2005.

[11] Z. H. Zhan *et al.*, "Adaptive Particle Swarm Optimization," *IEEE Trans. Sys. Man Cybern.*, Part B: Cybernetics, vol. 39, no. 6, 2009, pp. 1362–81.

[12] S. Scott-Hayward and E. Garcia-Palacios, "Utility-Based Resource Allocation for Real-Time IPTV in Wireless Networks," *Proc. IEEE Wireless Commun. Net. Conf.*, 2014.

BIOGRAPHIES

SANDRA SCOTT-HAYWARD [M'13] is a research fellow at the Institute of Electronics, Communications and Information Technology, Queen's University Belfast. She received a Ph.D. from Queen's University Belfast in 2013. She has been a Chartered Engineer since 2006, having worked as a systems engineer and engineering group leader with Airbus. Her research interests include wireless communications and networking, resource allocation, optimization and performance analysis for next generation networks, and security in software-defined networking.

EMILIANO GARCIA-PALACIOS is a lecturer at the Institute of Electronics, Communications and Information Technology, Queen's University Belfast. He received a Ph.D. from Queen's University Belfast in 2000 and since then has been leading research in wireless network resource management. His research interests include wireless protocols, wireless resource allocation, and optimization and traffic management for next generation gigabit networks.

AD HOC AND SENSOR NETWORKS

Edoardo Biagioni

Silvia Giordano

Recently, in this series, Conti *et al.* introduced the concept of the "Multi-Paradigm Era [1]," where several ad hoc paradigms may be mixed together and integrated with infrastructure-based networks. This issue fully confirms this trend. We present here three articles that clearly show how technologies and paradigms are not confined within a single field, but rather contribute in combination to provide useful solutions.

The first article we present in this issue deals with acoustic rangefinders, which is a promising technology for accurate proximity detection. In "Sparse Representation Based Acoustic Rangefinders: From Sensor Platforms to Mobile Devices," the authors survey state-of-the-art systems and identify possible challenges with acoustic rangers, including the number of available techniques and the processing power needed to derive useful information from raw ranging data. Solving such problems is fundamental for mobile computing and communications, in particular for sensor and mobile phone networks, which are the mainstream technologies for mobile applications. The authors further present an information processing framework based on sparse representation designed to address such challenges. This article concludes by introducing a low-energy acoustic ranging service for mobile devices called Beep-beep, and empirically evaluates its benefits.

The article by Li *et al.* discusses delay tolerant networks (DTNs) and how to test and evaluate them. DTNs are multi-paradigm in nature, as they rely on sensor and mobile phone networks, and furthermore can use multiple (wireless) technologies (WiFi, Bluetooth, etc.). In "Delay Tolerant Network Protocol Testing and Evaluation" the authors examine the evolution of DTN protocol testing and evaluation, and discuss the trend toward large-scale mobility trace supported emulation. They present a comprehensive survey of solutions for DTN simulation- and experiment-based testing, highlighting some limits in working with realistic scenarios. These authors propose a solution based on network virtualization, which is capable of efficiently handling the high complexity of realistic DTN scenarios. Such virtualization is able to simulate DTN environments reliably and gives an accurate system performance evaluation.

In "On the Potential of Bluetooth Low Energy Technology for Vehicular Applications," Lin *et al.* discuss a specific area of vehicular networks: the intra-vehicular wireless sensor network (IVWSN). At this time, such networks are considered central for improving the performance of vehicles, and are seen as a potential future support for innovative applications. The article focuses the potential usage of Bluetooth Low Energy (BLE) technology for bootstrapping IVWSNs and present an in-depth study to show BLE benefits for such networks.

The message we get from these articles is evident: the multi-paradigm era is a reality, and we are moving more and more toward different paradigms that are used for the same scenario, as in the case of DTN, and different scenarios that are solved with the same paradigm.

We also see that people are more and more at the center of each paradigm, both directly, with their mobility as in DTN or in mobile phone acoustic rangefinders, and indirectly, as in the case of vehicular networking.

As always, we want to give special thanks to all reviewers for their careful reviews of the articles and very helpful suggestions for improvement.

REFERENCES

[1] Conti *et al.* "Mobile Ad Hoc Networking: Milestones, Challenges, and New Research Directions," *IEEE Commun. Mag.*, Jan 2014.

BIOGRAPHIES

EDOARDO BIAGIONI (esb@hawaii.edu) is an associate professor in the Department of Information and Computer Sciences at the University of Hawaii at Manoa. His research interests focus on networking, with emphasis on ubiquitous wireless networking, but have over time ranged widely from security to high-performance computing, programming languages, and human-computer interfaces. He received his Ph.D. degree from the University of North Carolina at Chapel Hill, and has been a series co-editor for *IEEE Communications Magazine* since 2006.

SILVIA GIORDANO [M] (silvia.giordano@supsi.ch), Ph.D. from EPFL, is a full professor at SUPSI, Switzerland, and associate researcher at CNR. She directs the NetworkingLab. She has published extensively in the areas of QoS, traffic control, and wireless and mobile networking. She is co-editor of the books *Mobile Ad Hoc Networking* (IEEE-Wiley, 2004) and *Mobile Ad Hoc Networking: The Cutting Edge Directions* (Wiley 2013). She is an ACM Distinguished Scientist, ACM Distinguished Speaker, and on the Board of the ACM N2Women. She has been a series co-editor for *IEEE Communications Magazine* since 2004.

Sparse Representation Based Acoustic Rangefinders: From Sensor Platforms to Mobile Devices

Prasant Misra, Salil S. Kanhere, Sanjay Jha, and Wen Hu

Prasant Misra is with the Indian Institute of Science.

Salil S. Kanhere and Sanjay Jha are with the University of New South Wales.

Wen Hu is with CSIRO Computational Informatics.

ABSTRACT

Acoustic rangerfinders are a promising technology for accurate proximity detection, a critical requirement for many emerging mobile computing applications. While state-of-the-art systems deliver robust ranging performance, the computational intensiveness of their detection mechanism expedites the energy depletion of the associated devices that are typically powered by batteries. The contribution of this article is fourfold. First, it outlines the common factors that are important for ranging. Second, it presents a review of acoustic rangers and identifies their potential problems. Third, it explores the design of an information processing framework based on sparse representation that could potentially address existing challenges, especially for mobile devices. Finally, it presents μ-BeepBeep: a low energy acoustic ranging service for mobile devices, and empirically evaluates its benefits.

INTRODUCTION

The way people perceive, interpret, and use their personal space describes their interaction with other people and devices. *Distance* is an important proxemic cue that correlates the intimacy of such interactions with the respective physical span. Better user experiences (such as simultaneous photo sharing and "better-together" video viewing, device pairing, extrapolating human gesture, etc.) can therefore be delivered with a fine-grained distance knowledge of nearby users, or more specifically nearby users' devices. Essentially, (location-based) services that support such applications require high accuracy range information as an important prerequisite.

Motivated by the need for accurate proxemics, new ventures are taking the next logical step of developing the necessary technology using commodity software and hardware so that it can be readily used in commercial off-the-shelf (COTS) mobile devices such as cell phones, PDAs, notebooks, and so on. Of these portable devices, smartphones are particularly interesting because they are inevitably carried by people, and also have a wide range of sensors (such as microphone, camera, GPS, accelerometer, gyro, digital compass, and/or infra-red) and communication interfaces (such as 3G, GSM, WiFi, Bluetooth) for inferring proximity. When the proximity granularity is on the scale of a *few centimeters*, measuring the travel time of a sound pulse from the speaker (of one device) to the microphone (of another device) by counting the elapsed time samples (instead of timestamping) has been the most successful method [1]. The sample counting technique, though effective, is an expensive operation using a matched filter.[1] While smartphones have good computation and memory resources for its implementation, the longer execution time leads to greater energy usage that affects the *limited* battery life of these devices. This problem can therefore be simplified by designing a light-weight and faster mechanism that serves the purpose of sample counting.

In this regard it is interesting to note that the result of a matched filter has a *single* major spike at the correct time shift. The location of this data point is the only significant information needed for estimating the travel time of the sound pulse. A recent work by Misra *et al.* [2] exploited the sparse nature of the matched filter result for developing a detection method for low-cost and low-power sensor platforms. The key idea was to transfer the acoustic signal samples in a compressed form to a resourceful device to estimate the time delay without processing them locally on the device.

In this form of compression, the challenge is to preserve only the *sparse* information content that can later be correctly located. This problem is similar to the classic coin-puzzle where the task is to find the counterfeit coin (which is known to be heavier/lighter) among a group of identical coins (where all are genuine coins except one) using a balance scale with the fewest possible weightings. An efficient and quick solution is to weigh a selected group of coins (rather than a single coin) to correctly locate the fake coin.

Theoretical results in *sparse representation* are

[1] The matched filter is explained later.

0163-6804/15/$25.00 © 2015 IEEE

> *If a precise estimate of the distance between two devices is a key requirement, then previous studies have shown that the most successful techniques are based on measuring the travel time (also known as the time-of-flight (TOF), propagation delay, or time delay) of an acoustic signal propagation between them.*

also designed along the same principles as finding the odd coin. According to this theory, the signal can be compressed by measuring the summed intensity of random combinations of samples. If the (information) representation of the signal is sufficiently sparse in a given domain, then with a sufficient number of such measurements (but significantly fewer than the number of data samples), the recovery process can reproduce the correct result. This approach also facilitates the technique of split processing with cloud-offloading, which is an emerging trend to outsource data and computation load to remote servers and is a new paradigm to conserve energy[2] for devices with limited reserves. While such an approach is greatly useful for resource constrained platforms (such as wireless sensor nodes), its benefit on mobile devices needs investigation due to their advanced computation capability and power exhaustive wireless radio communication.

In this article we present a summary view of this field and demonstrate its benefits in conserving energy on mobile devices via our μ-BeepBeep system. Acoustic range-finding with μ-BeepBeep shows similar ranging performance as a standard system but at a (three times) lower energy cost, which is appreciable but not significant. It provides important lessons for building similar systems, and also serves as a demonstration tool for conducting research in this direction. To make this article self-contained, we present a review of ranging technologies and state-of-the-art acoustic ranging systems in the next section followed by an overview of μ-BeepBeep architecture and its performance. The final section reflects on the μ-BeepBeep ranging method and concludes with a summary of the article.

Ranging Technology

If a precise estimate of the distance between two devices is a key requirement, then previous studies have shown that the most successful techniques are based on measuring the *travel time* (also known as the time-of-flight (TOF), propagation delay, or time delay) of an acoustic signal propagation between them [4]. The reliability of this measurement depends on the *synchronization* accuracy of the local clocks (on each device) to a common time-scale, robustness of the *detection* method, and the design of the *ranging pulse*. We now revisit each of these basic principles.

Synchronization

Synchronization of time can be explicit or implicit. *Implicit* methods attain the state of synchronization as part of the ranging process without maintaining a separate service, as in *explicit* methods, to convert the time from one system clock to another. Time-difference-of-arrival (TDOA) based on hyperbolic theory and velocity-difference, round-trip time (RTT), and elapsed time between two time-of-arrivals (ETOA) are examples of implicit methods. With respect to acoustic ranging on sensor platforms and mobile devices, the methods of velocity-difference TDOA and ETOA have received particular attention.

Velocity-Difference TDOA (V-TDOA) — The main idea here is to transmit two synchronous signals with different speeds such as RF and sound, which traverse the same path length between a device pair. The receiver device detects the fast propagating RF pulse (almost instantaneously) followed by the slower sound pulse, and computes the separation distance by measuring the time delay between the arrival of these two signals.

Elapsed Time Between Two Time-of-Arrivals (ETOA) — It is a mathematical method to compensate the clock difference errors. ETOA,[3] introduced as part of the BeepBeep system [1], leads to precise timestamp recovery while allowing the local clocks to run asynchronously on the paired devices. Here the key idea is to use the difference in sample counts at each device instead of absolute sample counts.

The underlying technique is as follows. Let us say that there are two devices P and Q, each with their respective time-frame t_P and t_Q. A two-way reception is performed where device P emits a signal at t_{P0}, which is recorded by itself at time t_{P1}, and device Q at time t_{Q1}. Similarly, the same tasks are performed when device Q transmits the signal at t_{Q2}, which is recorded by itself at t_{Q3}, and device P at t_{P3}. Therefore, the received traces on each device should have two signals, one transmitted from the other device and one from itself. Next, the elapsed time between the arrival time of these two signal copies are exchanged with the other device. The difference of these two elapsed time measurements represent the two-way ranging time. Despite the arbitrary time delay between the two transmissions, the inter-device distance can be computed as:

$$\begin{aligned} d &= (1/2)(d_{P\to Q} + d_{Q\to P}) \\ &= (c/2)((t_{P3} - t_{P1}) - (t_{Q3} - t_{Q1})) + L'_P + L'_Q \quad (1) \end{aligned}$$

where, c is the speed of sound, and $L_P = 2L'_P$ and $L_Q = 2L'_Q$ are the distances between the transmitting and receiving sensors (i.e. speaker and microphone) on device A and B, respectively. It is important to note that such a mechanism can also work without self-recording if [t_{P0} and t_{P1}] and [t_{Q2} and t_{Q3}] can be made to converge at a single time value by using low-layer timestamping techniques (such as described in [6, 8] for resource constrained platforms, or in [9] for mobile phones).

Sensing and Detection

Detecting and obtaining an estimate of the arrival time of the ranging pulse can be performed by the following two approaches, which (characteristically) measure the received signal energy, but differ in their implementations and optimality criteria.

Energy Detector — Energy is measured by squaring and integrating the received signal over the observation interval. The integrator output is, subsequently, compared with a preset threshold to detect the presence of the signal, and the corresponding time taken to reach this level is the TOF estimate. While the detection mechanism is easy to implement, its performance is

[2] Energy savings offered by offloading computation is determined by three factors: wireless bandwidth, computation load, and transmitted data. Offloading is beneficial only when large amounts of computation are needed with relatively small amounts of communication [3].

[3] The basic device-to-device ETOA computation can be extended to a large-scale system consisting of many devices using the asymmetrical round trip based localization (ARTL) mechanism [5]. The ETOA is similar to RTT synchronization protocols such as TPSN [6] and TinySync [7], where the reference time of the devices in a networked system is set by the first transmission from the sending device, while the acknowledgment from the receiving device is used to mitigate detection and communication delays.

	System	Ranging technology	Signal design	Signal frequency	Signal time-period	Maximum range	Accuracy	Operational scope
Ultrasound	Active Bat (1999)	V-TDOA	Narrowband	—	—	—	3 cm	Indoor
	Cricket (2000)	V-TDOA	Narrowband	40 kHz	150 µs	10.50 m	2 cm	Indoor
	ALHoS (2001)	V-TDOA	Narrowband	40 kHz	—	3 m	2 cm	Indoor
	iBadge (2001)	V-TDOA	Narrowband	40 kHz	—	3 m	2 cm	Indoor
	Haza's sys. (2002)	V-TDOA	Broadband PN BPSK	[40-60] kHz	25 µs	3 m	< 1 cm	Indoor
	Whitehouse's sys. (2005)	V-TDOA	Narrowband	25 kHz	—	12 m	5 cm	Indoor
	WALRUS (2005)	V-TDOA	Narrowband	21 kHz	10 ms	10 m	Room level	Indoor
	M-Cricket (2011)	V-TDOA	Narrowband	40 kHz	150 µs	10.50 m	2 cm	Indoor
	Spiderbat (2011)	V-TDOA	Narrowband	40 kHz	—	14 m	< 1 cm	Indoor and outdoor
	TWEET (2011)	V-TDOA	Broadband	[20-25] kHz	50 ms	20 m	< 5 cm	Indoor and outdoor
Audible Acoustic	Calamari (2002)	TOF+RF-RSS	Narrowband	4.5 kHz	—	—	Node level	Indoor and outdoor
	Kwon's sys. (2005)	V-TDOA	Narrowband	4.3 kHz	—	20-30 m	33 cm	Outdoor
	Kushwaha's sys. (2005)	Message time-stamping	Broadband linear chirp	[0.05-5] kHz	—	30 m	[15-25] cm	Outdoor
	AENSBox (2006)	Message time-stamping	Broadband PN BPSK	[6-18] kHz	330 ms	[60-120] m	5 cm	Outdoor
	Thunder (2007)	FTSP+TOF	Narrowband	4.7 kHz	100 ms	137 m	1 m	Outdoor
	BeepBeep (2007)	ETOA	Broadband linear chirp	[2-6] kHz	50 ms	10 m	5 cm	Indoor and outdoor
	Whistle (2011)	Hyperbolic TDOA using ETOA	Broadband linear chirp	[2-6] kHz	50 ms	—	—	Indoor and outdoor
	EchoBeep (2012)	ETOA	—	—	—	10 m	10 cm	Indoor
	DeafBeep (2012)	Distance difference using ETOA	—	—	—	—	—	Indoor
	TWEET-v2 (2013)	V-TDOA	Broadband linear chirp	[1-20] kHz	10 ms	—	5 cm	(Semi) indoor

Table 1. Acoustic ranging systems.

dependent on the signal-to-noise ratio (SNR) of the received signal that is highly susceptible to uncertainty in noise power. Energy detection is optimal only if the noise power is known [10].

Matched Filter Detector — In this detection method the received signal is processed using a matched filter implemented by cross-correlating with the reference signal (i.e. a locally stored copy of the original transmitted signal). This operation results in correlation peaks, where the index of the first tallest correlation peak is the estimate of the pulse arrival time of the line-of-sight (LOS) path (that gives the range).

Matched filter detection is known to be optimal in stationary Gaussian noise conditions. However, in realistic situations it can be non-trivial to accurately estimate the LOS component due to the additive noise and multipath effects that obscure, or severely attenuate, the LOS (i.e. correlation) peak, or partially overlap and shift its index position in the correlation result [11]. Both of these conditions result in detection anomalies and timing errors.

Erroneous conditions can be controlled by broadening the bandwidth of the transmit signal, which translates to a narrower correlation peak for every multipath. This simple bandwidth adjustment provides greater temporal resolution, and hence alleviates the inaccuracies due to multiple propagation paths. This technique, referred to as pulse compression, not only resolves the different propagation paths but also increases the SNR of the LOS path without increasing the transmission power.

Erroneous conditions can be controlled by broadening the bandwidth of the transmit signal, which translates to a narrower correlation peak for every multipath. This simple bandwidth adjustment provides greater temporal resolution, and hence alleviates the inaccuracies due to multiple propagation paths.

Signal Design

The design of the ranging signal (i.e. bandwidth B and time-period T) plays a key role in delivering the desired ranging performance. Range resolution (and accuracy) depends on the bandwidth B of the ranging signal, while time period T is indicative of the signal energy (and hence, range).

Acoustic Systems: A Decade of Development

There has been more than a decade of development of in-air acoustic rangefinding systems, both in the *ultrasound* (> 20 kHz) and *audible acoustics* (< 20 kHz) band [4]. Table 1 provides a comparative summary of the existing acoustic systems. However, within the scope of this article we will focus on audible acoustics rangefinders.

Calamari and the system designed by Kwon *et al.* are the early rangefinders in this domain. They are based on early WSN platforms (such as Mica2), and use COTS piezo-electric buzzers to generate (short) ranging pulses that are sinusoidals of ≈ 4.5 kHz. Due to their limited (narrowband)[4] frequency span, they are highly sensitive to environmental noise and are also incompetent of resolving the energy in the multipath echoes. Kwon *et al.*'s system obtained an accuracy of ≈ 33 cm (over a distance of 20–30 m), while Calamari only provided node-level ranging resolution. Thunder also used a similar signal design, but instead increased the ranging signal length to 100 ms to obtain better range, which translated to a range improvement of 137 m with an average accuracy of 1 m.

To overcome the range and accuracy/resolution limitations (as reported by Calamari, Kwon's system, and Thunder), the following generation of rangefinders such as the system by Kushwaha *et al.* AENSBox, Thunder, BeepBeep, Whistle, EchoBeep, DeafBeep, and TWEET-v2, introduced various design changes. They share a common detection technique based on matched filter, but they differ in their signal designs, synchronization schemes, and methods to improve the received SNR.

Kushwaha *et al.*'s system was based on the Mica2 platform with an attached custom 50 MHz DSP processor and an external speaker. The ranging signal was a Gaussian windowed linear chirp of 0.05–5 kHz. It employed a message time stamping technique for synchronizing, and added a series of consecutive position-modulated chirps at the same phase and averaged these measurements to enhance the SNR of the received signal.

AENSBox was comprised of a custom designed acoustic sensor array that utilized beamforming to improve the received SNR, and a separate time synchronization services (based on low-level time stamping) to prevent clock skew and drifting. The ranging signal was a 2048-chip code modulated using binary phase shift keying (BPSK) on a 12 kHz carrier spread over 6–18 kHz. It differed from most of its predecessors (including ultrasonic systems) in the use of a separate synchronization service that maintained metrics to convert from one system clock to another on demand, rather than a synchronous radio and audio pulse. This approach is beneficial in scenarios where the audio range is greater than the radio range.

BeepBeep was a software-based ad hoc infrastructure-less solution implemented on COTS mobile devices rather than on non-standard platforms. The ranging pulse was a 50 ms linear chirp of 2-6 kHz, and used ETOA to avoid explicit clock synchronization. The techniques of EchoBeep, DeafBeep, and Whistle improve on the basic ranging technique demonstrated by BeepBeep. While EchoBeep and DeafBeep, respectively, are provisioned for NLOS conditions and distance difference measurements (for devices without microphones), Whistle measures the TDOA and is based on peak detection. Whistle detects the arrival time of the ranging pulse by the sequential change-point mechanism, where the key idea is to identify the first arriving signal that deviates from the noise after filtering out background noise. This technique is useful when there is relatively little noise and interference from outside sources in the frequency range of the ranging pulse (such as inside a car), and therefore cannot be used for general ranging applications.

TWEET-v2, while adhering to best design practices of existing acoustic rangefinders (especially AENSBox), introduced two additional features. First, it adopted an envelope (instead of peak) detection technique to make the role of sidelobes of the cross-correlation function (of the matched filter) irrelevant, and counter the effect of noise through the least-square curve fitting approach. Such an approach also provided the

[4] Signals can be categorized into narrowband and broadband depending on their relative bandwidth (i.e., the ratio of the signal bandwidth to the center frequency). Narrowband signals have a relative bandwidth of less than 0.2, while the same exceeds 0.2 for broadband signals.

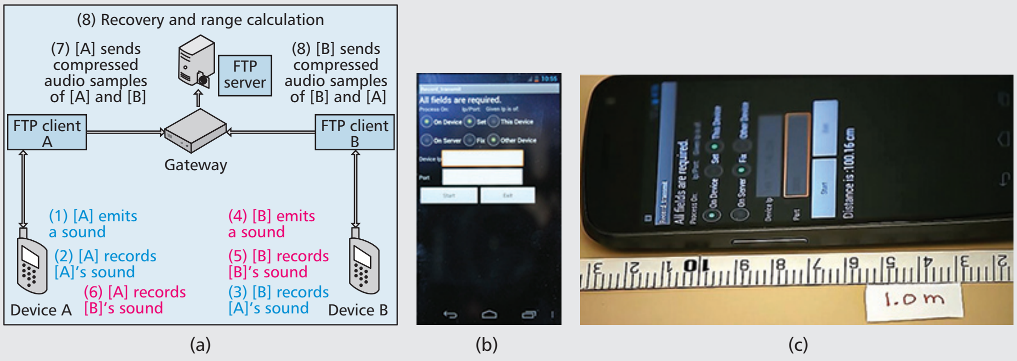

(8) Recovery and range calculation
(7) [A] sends compressed audio samples of [A] and [B]
FTP server
(8) [B] sends compressed audio samples of [B] and [A]
FTP client A
FTP client B
Gateway
(1) [A] emits a sound
(4) [B] emits a sound
(2) [A] records [A]'s sound
(5) [B] records [B]'s sound
(6) [A] records [B]'s sound
(3) [B] records [A]'s sound
Device A
Device B
All fields are required.
Distance is :100.16 cm
1.0 m
(a)
(b)
(c)

Figure 1. The design and implementation of μ-BeepBeep: a) system architecture; b) and c) user interface on the mobile device.

benefit of finer resolution that can be fractions of a sampling period. However, its performance can be limited by the SNR condition, where the proposed interpolation method will introduce significant bias under low SNR. Second, it introduced a near-inaudible ranging signal for mutual coexistence with existing acoustic signals in the audible domain. On the performance side, TWEET-v2 showed similar ranging statistics as AENSBox except for improvements in power consumption, processing cost, and (minor) accuracy.

Audible acoustic systems that use a broadband signal design and measure the TOF with a matched filter have reported impressive performance in terms of accuracy, range, and coverage. Although processing acoustic signals in the audible range (below 20 kHz) requires sampling rates of less than 50 kHz, resource constrained sensor platforms are still unable to perform on-board processing, and require additional hardware components (such as DSP or other specialized processors and units) that are power exhaustive and costly. An attempt to simplify this problem was first demonstrated by Misra *et al.* [2]. Here the key idea was to compress (rather than process) the received signal samples and transmit the condensed data to a more resourceful device (or base-station) that can estimate the range by using the techniques of *sparse approximation*. Therefore, outsourcing data and computation to a remote computing platform (also known as cloud-offloading) reduced the local burden on the device. In the following section we investigate the benefits of such a mechanism to conserve the battery life of mobile devices.

Although processing acoustic signals in the audible range (below 20 kHz) requires sampling rates of less than 50 kHz, resource constrained sensor platforms are still unable to perform on-board processing, and require additional hardware components that are power exhaustive and costly.

Acoustic Ranging with μ-BeepBeep

We now present μ-BeepBeep, a low energy acoustic ranging service for mobile phones. In this section we discuss the theory of sparse approximation, and then adopt its working principle for ranging as part of μ-BeepBeep.

An Overview of Sparse Approximation

Motivating Insight — One can accurately and efficiently recover the information of a high dimensional signal (say **x**) from only a small number of compressed measurements, when the signal-of-interest is sufficiently *sparse* in a certain transform domain [12, 13].

The sparsifying domain, referred to as a *dictionary* $\Psi \in \mathbb{R}^{n \times d}$, is a collection of parametrized waveforms that expresses **x** as a linear combination of a few significant elements. It is represented as: $\mathbf{x} = \Psi\mathbf{s} = \Sigma_{i=1}^{d} s_i\psi_i$ where $\mathbf{s} \in \mathbb{R}^d$ is a coefficient vector of **x** in the Ψ domain, and ψ_i is a column of Ψ. If **s** is sparse,[5] then it is possible to recover the position and value of its coefficients by a combinatorial problem, but is intractable. In pursuit of a polynomial time solution, Donoho [12] showed that, for a large system of equations, **s** can still be recovered by the following ℓ^1-minimization problem with high probability.

$$(\ell^1): \hat{\mathbf{s}}1 = \arg\min \|\mathbf{s}\|_1 \quad \text{subject to: } \mathbf{x} = \Psi\mathbf{s} \qquad (2)$$

It has also been shown that dimensionality *reduction* by *random* linear projections preserves the ℓ^2 distance (i.e. all useful information) in the projection domain [14]. ℓ^1-minimization can still be used to recover the sparse **s** from the projected measurements with an overwhelming probability, even though its dimension is significantly reduced. This operation can be achieved using a random sensing matrix $\Phi \in \mathbb{R}^{m \times n}$ as: $\mathbf{y} = \Phi\mathbf{x} = \Phi(\Psi\mathbf{s}) = (\Phi\Psi)\mathbf{s}$, where $m \ll n$ and $\mathbf{y} \in \mathbb{R}^m$ is the measurement vector. However, for recovery the columns of (ΦΨ) should be as independent as possible so that the information regarding each coefficient of **s** is contributed by a different direction, and this is achievable if Φ and Ψ are *more* incoherent.[6] Ensembles of random matrices sampled independently and identically (i.i.d.) from Gaussian and ± 1 Bernoulli distributions are largely incoherent with any fixed dictionary Ψ, and therefore permit computationally tractable recovery of **s** [12, 13].

Road-Map — The critical aspect for casting the acoustic ranging problem into the general framework of sparse approximation is to design a sparsifying representation dictionary. We use a Toeplitz matrix (an inverse of the Hankel matrix[2]) as the sparsifying dictionary in the design of μ-BeepBeep, and demonstrate its ranging performance.

[5] *Sparsity*: It characterizes the *signal-of-interest*. It expresses the signal as a linear combination of few elements taken in a particular dictionary as: $\mathbf{x} = \Psi\mathbf{s}$. The *dictionary* Ψ transform the time-domain representation of **x** that is, typically, dense to its equivalent representation s that is, typically, sparse.

[6] *Incoherence*: It pertains to the *sensing modality* (ΦΨ). Its purpose is to measure the unknown signal components and store this information in **y**. Therefore, **y** is the result of a combination of the columns of (ΦΨ), each one weighted by a different component of **s**. For recovery, the columns of (ΦΨ) should be as *independent* as possible; and this is achievable if Φ and Ψ are more incoherent. Incoherence also makes the process of compression (using Φ) completely independent of the choice of the sparsifying dictionary Ψ.

This result suggests that information embedding in random ensembles preserves the energy of its respective higher-dimension representation, as opposed to deterministically choosing samples and discarding information by downsampling.

System Architecture

Figure 1 shows the system architecture of μ-BeepBeep [15] with only two mobile devices (P and Q). The basic ranging technique is similar to BeepBeep wherein a two-way sensing of audio signals is performed as per the ETOA mechanism. Unlike BeepBeep, where each device estimates the TOA (locally using a matched-filter) and exchanges this information over a WiFi[7] radio link to compute the distance, μ-BeepBeep takes a different approach. Here, instead of locally processing the recorded audio samples, each device compresses the raw data and transfers it to a back-end server that estimates the range using the sparse approximation framework.

The μ-BeepBeep protocol consists of the following three steps:

- **Initiation**: The process is started by an initiating device, which disseminates a ranging schedule in an initiation message to all the participating devices (in our case, it is only 1).
- **Ranging**: Each pair of devices performs two-way acoustic ranging as per the ETOA mechanism using a linear chirp [2-6] kHz/0.05 s.
- **Detection and post-processing**: It is performed in two phases. In the following, we explain it with respect to a single device as it is identical for all devices involved in ranging.

Compression — At each device, the dimension of the received audio signal $\mathbf{x}$ is significantly reduced by multiplying it with a random sensing matrix $\Phi \in \mathbb{R}^{m \times n}\}$ resulting in the measurement vector $\mathbf{y} \in \mathbb{R}^m (m \ll n)$ as: $\mathbf{y} = \Phi x$. m is related to n by the compression factor α given as: $m = \alpha\, n$, where $\alpha \in [0, 1]$. Φ is a binary sensing matrix with its entries i.i.d. sampled from a balanced symmetric Bernoulli distribution of ± 1. A balanced Φ consists of ± 1 at equal probability, where each row contains an equal number of 1's and –1's. Therefore, in each row of Φ, the sum of the elements is always zero. A balanced Φ provides a higher probability of detection (at recovery) if the noise in $\mathbf{x}$ is Gaussian. The m samples of $\mathbf{y}$ are transferred to the remote server using the file transfer protocol (FTP) service.

Reconstruction and Detection — The server requires the a-priori knowledge of the seed that generates Φ and the dictionary Ψ [2]. The ranging framework recovers the sparse correlation coefficient vector $\mathbf{s}$ by solving the following Eq. 3 (a modified version of Eq. 2), which is more resistant to measurement noise with a given tolerance ε.

$$(\ell_r^1)\text{: } \hat{\mathbf{s}} = \arg\min \|\mathbf{s}\|_{\ell_1} \quad \text{s.t.: } \|\Phi\Psi\mathbf{s} - \mathbf{y}\|_2 \le \varepsilon \qquad (3)$$

Ψ is the positive and negative time shifted Toeplitz matrix constructed from the known samples of the transmitted signal $\mathbf{x}$. $\hat{\mathbf{s}}$ is related to the various propagation paths between the transmitting and receiving devices, where the index of the first tallest correlation coefficient peak is the TOA estimate.

Distance Calculation — The server recovers the two TOAs for each device pair, and computes the ETOA according to Eq. 1 to estimate the separation distance. In Eq. 1 the distance $L'_P + L'_Q$ is a constant to the respective device, and is measured a priori.

System Implementation and Study

We implemented the μ-BeepBeep ranging service on Google Nexus, a COTS smart phone. The device was running Android version 4.0. It featured a 1.2 GHz dual-core ARM Cortex-A9 processor, 1 GB RAM, and a wide range of sensors and communication interfaces. Within the scope of this project, we only used the speakers, microphones, and WiFi radio of the device.

The study was conducted in an indoor and outdoor setup. Here we only present the results for the indoor environment, which was a lecture theatre of dimension [25 × 15 × 20] m. Figures 2a and 2b explain the notion of sparsity, wherein the received signal in the time domain has a sparse representation (only two samples have dominant peaks among others that take on zero or negligible values) in the transform (i.e. correlation) domain. Figures 3a and 3b, respectively, show the detection result using a matched filter and sparse approximation. We observe that both methods obtain exactly the same estimate for the position of the detection peaks along with the remaining multipath profile.

A key decision is to choose the optimal compression factor α that achieves the best accuracy with the least measurements m as a smaller m leads to lower storage and transmission cost. m depends on the sparsity of the received signal in the correlation domain, which in turn depends on the received SNR that varies with transmission power and ranging distance. We performed a series of experiments to empirically estimate m. In this setup the two mobile devices were fixed at a constant separation distance of 3 m. The (acoustic) transmit power was varied such that the received SNR were recorded within the limits: [0–5) dB, [5–10) dB, [10–20) dB, [20–30) dB. Figure 4a shows the dependence of a-compression and its recovery accuracy on the SNR. The figure shows the relative mean error and its deviation with respect to the (best-case) of using a matched filter. The errors are as large as 1 m with α = 0.05, but attain stability at α = 0.30. This method of sparse approximation shows an order of magnitude 2 improvement in detection accuracy over the deterministic downsampling technique (Fig. 4b). This result suggests that information embedding in random ensembles preserves the energy of its respective higher-dimension representation, as opposed to deterministically choosing samples and discarding information by downsampling.

In Figs. 4c and 4d we compare the ranging performance and energy consumption profile of four different ranging techniques:[8]

- BeepBeep (sense, xcorr, wireless data exchange between the ranging devices).
- μ-BeepBeep (sense, compress, wireless data transfer to server, sparse approximation).
- s-BeepBeep (sense, wireless data transfer to server, sparse approximation).
- e-BeepBeep (sense, energy detection, wireless data exchange between the ranging devices).

[7] WiFi, due to its high wakeup and connection maintenance time, may not be the best choice for ranging applications (operating in a standalone manner) that emphasize on minimizing energy consumption. For our set of applications, we assume that our ranging technique will be used as a service that piggybacks on existing WiFi connections; which, will prove to be more economical (in terms of energy) than using it as a standalone decoupled module.

[8] All four different systems were implemented on the same platform as described previously.

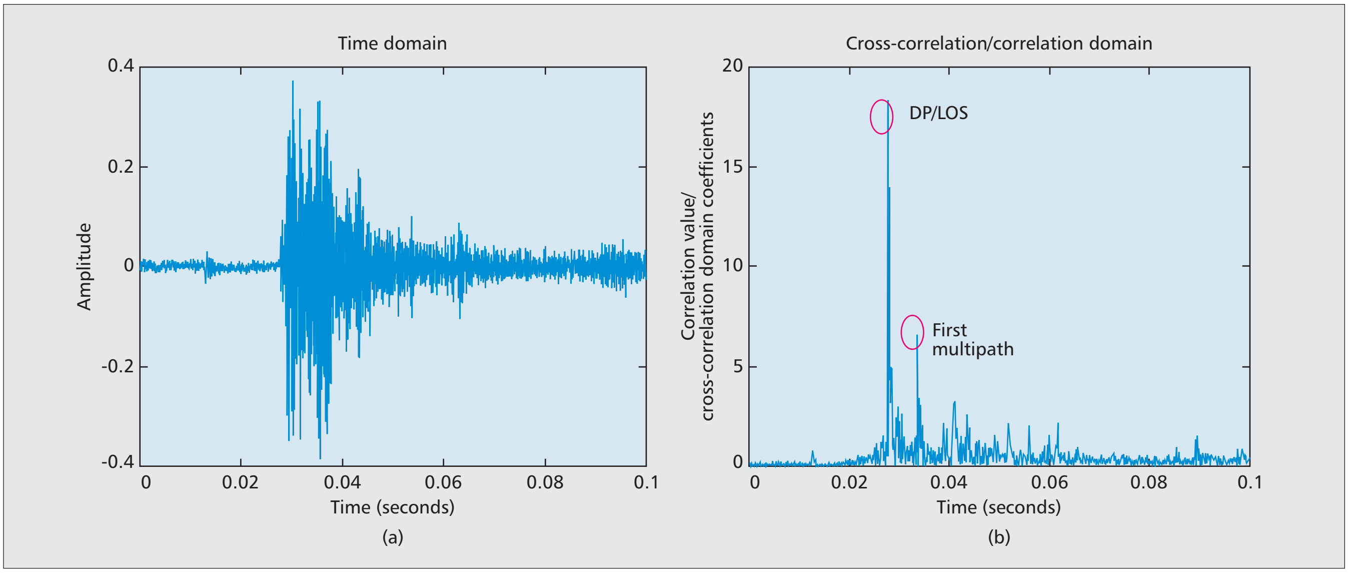


Figure 2. Equivalent representations of the same waveform in: a) time domain and b) correlation domain.

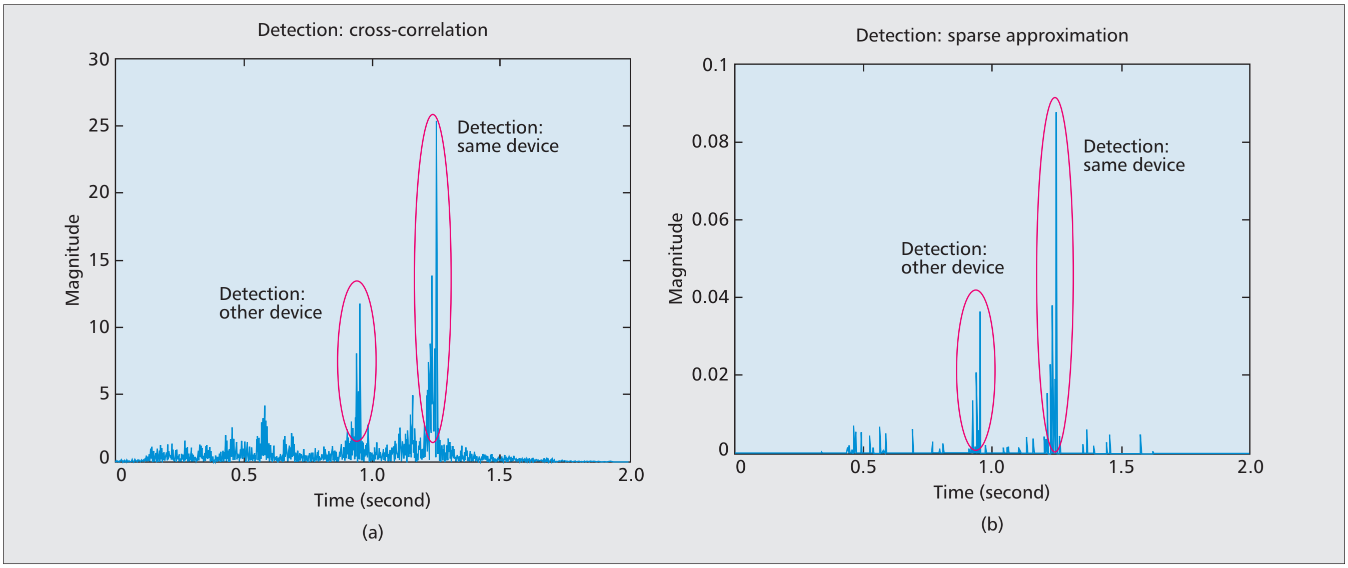


Figure 3. The LOS peak is correctly detected by both the methods: a) standard cross-correlation and b) sparse approximation.

Device *P* was fixed while the device *Q* was moved along the direct LOS in a controlled manner. The correct ground truth was established using a measuring tape and markers. The speed of sound used in distance calculation was according to the model: c_{air} =331.3 + 0.6θ (θ: air temperature in °C).

All systems achieved a maximum ranging distance of 8 m, except e-BeepBeep. While it has the lowest energy profile, the ranging performance (even for distances < 4 m) is highly erroneous that stems from its inferior noise power compensation capability. μ-BeepBeep recorded almost similar performance as BeepBeep for [1–4] m. Thereafter, its performance deteriorated for distance measurements from [5–8] m, which was still within 2 cm of the respective measurement error reported by BeepBeep. These lower accuracy measurements of μ-BeepBeep were due to the decrease in signal sparsity (an important factor for efficient reconstruction) with lower SNR of the received signals. In addition, μ-BeepBeep is more than 2.5 times energy efficient than BeepBeep (Fig. 4d). s-BeepBeep is also able to mirror the performance of BeepBeep as the underlying sparse approximation performed on uncompressed data is similar to cross-correlation, and also has energy benefits.

Discussion and Conclusion

The sparse representation based acoustic ranging technique can efficiently capture and embed information in a lower dimensional space (by random ensembles), and subsequently recover it from an underdetermined system. Such an approach has several merits. First, it provides a simple dimensionality reduction mechanism to condense the dataset. As the data compressibility is proportional to its information level, sparse (information) signals can be compressed significantly. Second, it requires transferring and pro-

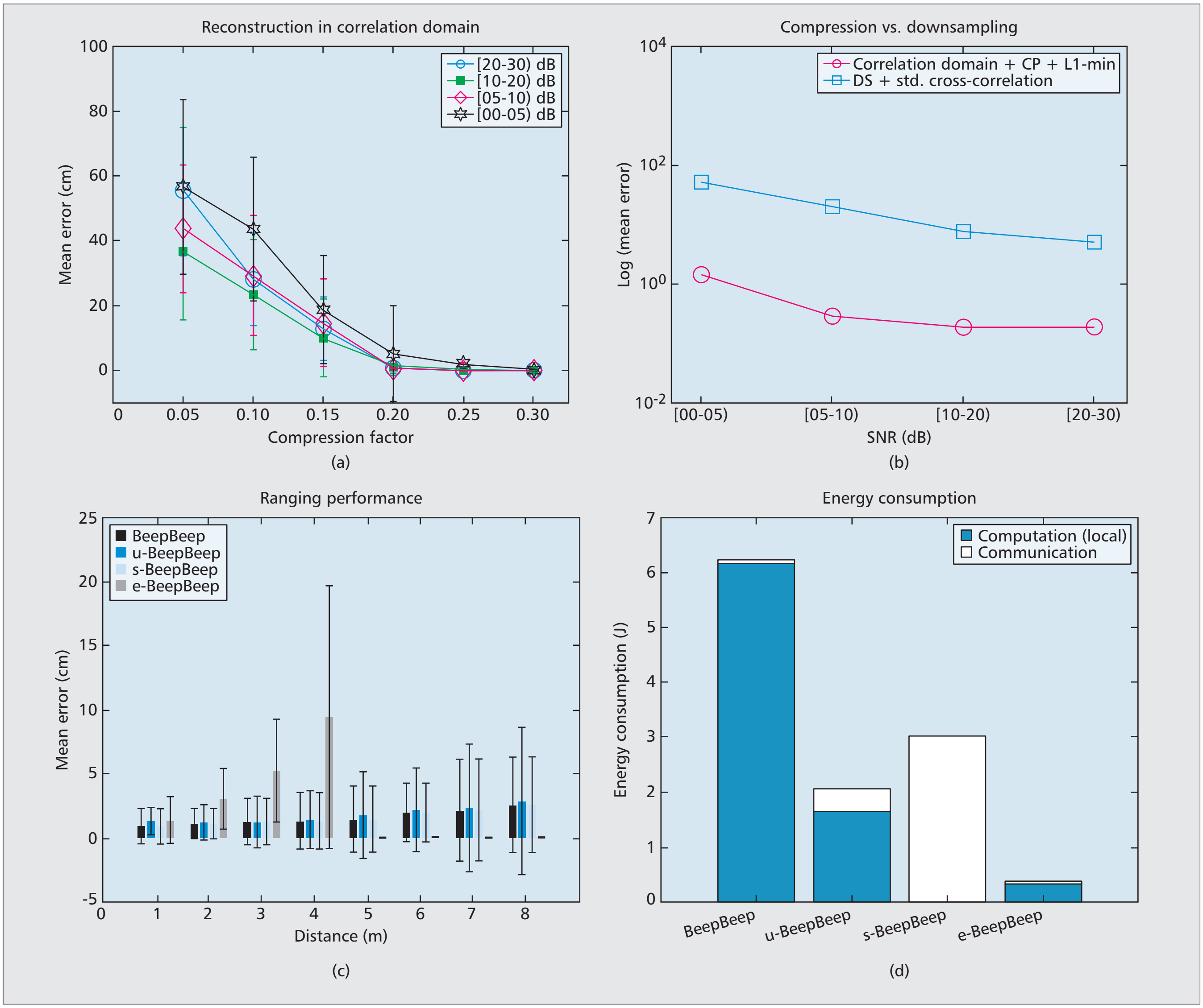


Figure 4. Empirical results in the indoor setup: a) characterization of compression factor α with SNR; b) for a compression/downsampling factor of 0.30, the sparse approximation method based on l1-min in the correlation domain shows an order of magnitude 2 higher detection accuracy compared to downsampling with cross-correlation; c) and d) ranging performance and energy profile of: (1) BeepBeep (sense, xcorr, wireless data exchange between the ranging devices), (2) µ-BeepBeep (sense, compress, wireless data transfer to server, sparse approx.), (3) s-BeepBeep (sense, wireless data transfer to server, sparse approx.), and (4) e-BeepBeep (sense, energy detection, wireless data exchange between the ranging devices).

cessing a significantly smaller dataset to obtain accuracies comparable to the state-of-the-art detection technique. At the local device end, the simplicity of this operation translates into appreciable energy savings. Third, it introduces a system controlled parameter for accuracy moderation (α), wherein applications with variable position uncertainty can use this feature to further optimize the energy budget of the device. For example, applications that require lower-range accuracy (e.g. 100 cm) could use only five percent (instead of 30 percent) measurements (Fig. 4a). Unlike traditional techniques, it is asymmetric wherein the two entities (i.e. mobile device and remote server) need not know the compression basis. This forms the basis for secure ranging, which could further be developed along the lines of asymmetric key cryptography. Finally, in comparison to lossless compression techniques (such as LZ77-based algorithm "gzip"), it offers a graceful performance degradation in the event of packet loss during data transmission as it can still recover the results, but with larger errors, and has the same performance as compressing with a smaller α.

While it is an important milestone to be able to reduce the energy demands on the device, the next hurdle is in the recovery process that incurs mammoth computation loads that require longer execution time. Such a practice would exacerbate the loads on servers (and from a cloud perspective, data centers and energy grids) that are already under pressure from our mobile demands.

In conclusion, this article outlines features common to ranging, with a special focus on acoustic techniques, and then summarizes the existing acoustic systems developed over the last decade. It identifies those factors that present the greatest challenge to energy depletion. Finally, the article discusses the emerging technology of sparse representation-based acoustic rangefinders as an important step toward closing the gap between performance and energy consumption, and provides empirical evidence using the µ-BeepBeep system.

Acknowledgment

We acknowledge Gurmanjeet S. Sidhu and Arundeep Kamboj (Indian Institute of Technology-Patna, India) for their help with implementation of µ-BeepBeep. This work was supported in part by the Australian Research Council Discovery Grant DP110104344.

References

[1] C. Peng *et al.*, "BeepBeep: A High Accuracy Acoustic Ranging System Using COTS Mobile Devices," *SenSys '07*, New York, NY, USA, 2007, ACM, pp. 1–14.

[2] P. Misra *et al.*, "Efficient Cross-Correlation via Sparse Representation in Sensor Networks," *IPSN '12*, New York, NY, USA, 2012, ACM, pp. 13–24.

[3] K. Kumar and Y.-H. Lu, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?" *Computer*, vol. 43, no. 4, April 2010, pp. 51–56.

[4] P. Misra *et al.*, "Acoustical Ranging Techniques in Embedded Wireless Sensor Networked Devices," *ACM Trans. Sensor Networks*, vol. 10, no. 1, Dec. 2013, pp. 15:1–15:38.

[5] L. Bin *et al.*, "Asymmetrical Round Trip Based Synchronization-Free Localization in Large-Scale Underwater Sensor Networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, 2010, pp. 3532–42.

[6] S. Ganeriwal, R. Kumar, and M. B. Srivastava, "Timing-Sync Protocol for Sensor Networks," *SenSys '03*, New York, NY, USA, 2003, ACM, pp. 138–49.

[7] M. L. Sichitiu and C. Veerarittiphan, "Simple, Accurate Time Synchronization for Wireless Sensor Networks," vol. 2 of *WCNC '03*, IEEE, 2003, pp. 1266–73.

[8] M. Maróti *et al.*, "The Flooding Time Synchronization Protocol," *SenSys '04*, New York, NY, USA, 2004. ACM, pp. 39–49.

[9] S. Sur, T. Wei, and X. Zhang, "Autodirective Audio Capturing Through a Synchronized Smartphone Array," *MobiSys '14*, New York, NY, USA, 2014, ACM, pp. 28–41.

[10] F. F. Digham, M. S. Alouini, and M. K. Simon, "On the Energy Detection of Unknown Signals over Fading Channels," *IEEE Trans. Commun.*, vol. 55, no. 1, 2007, pp. 21–24.

[11] K. Pahlavan, L. Xinrong, and J. P. Makela, "Indoor Geolocation Science and Technology," *IEEE Commun. Mag.*, vol. 40, no. 2, 2002, pp. 112–18.

[12] D. L. Donoho, "For Most Large Underdetermined Systems of Linear Equations the Minimal l1-Norm Solution is also the Sparsest Solution," *Commun. on Pure and Applied Mathematics*, vol. 59, no. 6, 2006, pp. 797–829.

[13] E. J. Candes and T. Tao, "Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, 2006, pp. 5406–25.

[14] R. Baraniuk *et al.*, "A Simple Proof of the Restricted Isometry Property for Random Matrices," *Constructive Approximation*, vol. 28, 2008, pp. 253–63.

[15] G. S. Sidhu *et al.*, Poster Abstract: m-BeepBeep: Low Energy Acoustic Ranging on Mobile Devices, *EWSN '13*, 2013.

Biographies

PRASANT MISRA is a senior member, technical staff at the Robert Bosch Centre for Cyber Physical Systems in the Indian Institute of Science, Bangalore. He received his Ph.D. from the University of New South Wales, Sydney in 2012. His current research interests include low-power sensing/communication and energy-efficient computing with a focus on system design and implementation within the general framework of cyber physical systems and Internet of Things. He has many years of experience in technology development, and has worked in different roles and capabilities for Keane Inc. (now a unit of NTT Data Corporation), India; CSIRO ICT Centre, Australia; Red Lotus Technologies, USA; and SICS Swedish ICT, Sweden. The outcome of this work has either resulted in commercial products, or publications in premier sensornet forums such as ACM/IEEE IPSN and ACM TOSN. His professional and research contributions have been recognized by numerous awards, of which it is noteworthy to mention the ERCIM Alain Bensoussan/Marie Curie Fellowship (2012) and the AusAID Australia Awards Leadership Program (2008). He has served on the organizing/technical committee of a number of international conferences/events. Currently, he serves as an associate technical editor of IEEE Communication Magazine, member of TiE (Bangalore) SIG IoT, ACM and IEEE."

SALIL KANHERE received his M.S. and Ph.D. degrees, both in electrical engineering, from Drexel University, Philadelphia in 2001 and 2003, respectively. He is currently an associate professor in the School of Computer Science and Engineering at the University of New South Wales in Sydney, Australia. His current research interests include embedded sensor networks, pervasive computing, participatory sensing, mobile networking, privacy, and security. He has published over 125 peer-reviewed articles and delivered over 15 tutorials and keynote talks on these research topics. Salil was the program co-chair for IEEE LCN 2014 and IEEE MoWNet 2014, and has also served on the organizing committee of a number of IEEE and ACM international conferences (e.g. ACM SenSys, ACM CoNext, IEEE WoWMoM, IEEE LCN, ACM MSWiM, IEEE SenseApp, and ISSNIP). He currently serves as the area editor for pervasive and mobile computing for the *International Journal of Ad Hoc and Ubiquitous Computing* and the *ICST Journal on Ubiquitous Environments*. Salil is a senior member of both the IEEE and the ACM and also serves on the IEEE CS TCCC Executive Committee. Salil is a recipient of the Alexander von Humboldt Research Fellowship in 2014.

SANJAY K. JHA is a professor and head of the Network Group at the School of Computer Science and Engineering at the University of New South Wales. His research activities cover a wide range of topics in networking including network and systems security, wireless sensor networks, adhoc/community wireless networks, resilience and multicasting in IP networks. Sanjay has published over 160 articles in high quality journals and conferences. He is the principal author of the book *Engineering Internet QoS* and a co-editor of the book *Wireless Sensor Networks: A Systems Perspective*. He served as an associate editor of the *IEEE Transactions on Mobile Computing* (TMC) and was on the editorial board of *ACM Computer Communication Review* (CCR).

WEN HU is a principal research scientist and research project leader at CSIRO Digital Productivity Flagship. Much of his research career has focused on the novel applications, low-power communications and security issues in sensor networks and pervasive computing systems. He has recently become interested in the applications of compressive sensing in Internet of Things (IoT). Hu has published regularly in the top rated sensor network venues and mobile computing such as ACM/IEEE IPSN, ACM SenSys, EWSN, *ACM Transactions on Sensor Networks* (TOSN), *Proceedings of the IEEE*, and *Ad-hoc Networks*. Hu received his Ph.D from the University of New South Wales (UNSW) in computer science and engineering. He is a recipient of prestigious CSIRO Office of Chief Executive (OCE) Julius Career Award (2012-2015) and CSIRO OCE postdoctoral grant. Hu holds adjunct associate professor positions at Queensland University of Technologies, University of Queensland, and UNSW. He is a senior member of ACM and IEEE, and served regularly on the organizing and program committees of networking conferences including ACM/IEEE IPSN, ACM SenSys, ACM MobiSys, IEEE LCN, IEEE ICC, IEEE WCNC, IEEE DCOSS, IEEE GlobeCom, IEEE PIMRC, and IEEE VTC.

> *While it is an important milestone to be able to reduce the energy demands on the device, the next hurdle is in the recovery process that incurs mammoth computation loads that require longer execution time. Such a practice would exacerbate the loads on servers that are already under pressure from our mobile demands.*

Delay-Tolerant Network Protocol Testing and Evaluation

Yong Li, Pan Hui, Depeng Jin, and Sheng Chen

Yong Li and Depeng Jin are with Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University.

Pan Hui is with Hong Kong University of Science and Technology, Telekom Innovation Laboratories, and Aalto University.

Sheng Chen is with the University of Southampton and King Abdulaziz University.

This work is supported by National Basic Research Program of China (973 Program) (No. 2013CB329105), National Nature Science Foundation of China (No. 61301080, No. 91338102, 61321061, and No. 61171065), National High Technology Research and Development Program (No. 2013AA013501 and No. 2013AA013505), and China's Next Generation Internet (No. CNGI-12-03-007).

ABSTRACT

Delay-tolerant networks, DTNs, are characterized by lacking end-to-end paths between communication sources and destinations. A variety of routing schemes have been proposed to provide communication services in DTNs, and credible and flexible protocol evaluation tools are in demand in order to test these DTN routing schemes. By examining the evolution of DTN protocol testing and evaluation, this article discusses the trend toward large-scale mobility trace supported emulation, and we propose TUNIE, a large-scale emulation testbed for DTN protocol evaluation based on network virtualization. Unlike the existing simulation tools and real-life testbeds, which either cannot provide a realistic DTN environment setup or are too costly and time-consuming, our proposed TUNIE architecture is capable of simulating reliable DTN environments and obtaining an accurate system performance evaluation. By system prototype and implementation, we demonstrate TUNIE as a flexible platform for evaluating DTN protocol performance.

INTRODUCTION

Delay-tolerant networks (DTNs) [1, 2] have attracted lots of attention in the past 10 years, and many related interesting applications have been experimented and tested, including mobile social networks based on human mobility, sensor networks for wildlife tracking and habitat monitoring, vehicular ad hoc networks for road safety and commercial applications, and deep-space interplanetary networks. In a DTN, most of the time there are no end-to-end paths from communication sources to destinations due to node mobility, wireless propagation effects, sparse node density, and other adverse factors. For this kind of network, traditional ad hoc routing protocols, which rely on end-to-end paths, fail to work [1]. Therefore, a new routing mechanism, called store-carry-and-forward [3], was proposed to provide communication. In order to improve message delivery probability, a variety of routing schemes have been proposed, such as two-hop relaying, spray and wait, and MaxProp [3], which aim to reduce the overhead of epidemic routing. Furthermore, some of these routing schemes claim to obtain optimal system performance, and typically they attempt to achieve short message delivery delay with relatively low transmission cost. However, there is a trade-off between message delivery delay and delivery cost. Generally speaking, shorter delivery delay is obtained at the expense of higher cost, and vice versa. Therefore, it is critically important to accurately evaluate these routing schemes in order to show their advantages and drawbacks objectively.

Recently, theoretical analysis frameworks, such as Markov models [4] and ordinary differential equation (ODE) models [5], are being used to evaluate the performance of DTN protocols. However, these models are far too simplified to be capable of faithfully representing highly complicated DTNs, and the ability of these models to evaluate DTN protocols is severely limited. Therefore, more realistic simulation and experiment-based evaluation tools are needed [3, 6]. Current testing tools can be classified into two types: software-based simulation and testbed-based experimentation. Software-based simulation can be carried out using general network simulation software like NS-2 and OPNET, or specialized DTN simulation tools like ONE [7] and OMNeT++ [8]. Recently emerging real-life DTN testbeds include UMass DieselNet [9] and ORBIT [10], which are built to carry out experiments for evaluating DTN related algorithms and protocols. These testbeds offer realistic DTN environments. However, setting up an experiment in such a testbed involves vast investment in terms of money and time.

In this article, we first review the evolution of the protocol testing and evaluation for DTN and discuss the trend toward large-scale mobility trace supported emulation by surveying the emerging approaches to DTN protocol testing. In order to overcome the shortcomings of both the existing simulation tools and experimental testbeds, we propose TUNIE, a large-scale emulation testbed for DTN protocol evaluation based on network virtualization, which offers the following highly desired features. First, TUNIE enables the implementation of realistic environments for credible evaluation of DTNs by controlling the data transmission through regulated wired and wireless links. Second, it provides deep programmability of networking functions to customize system-level parameters as well as

0163-6804/15/$25.00 © 2015 IEEE

Categories	Main features	Representative platforms
General simulation software	Software simulated networks, simple mobility models	NS-2, NS-3, OPNET
DTN specific platform	Discrete event simulator, realistic mobility trace integration	OMNeT++[8], The ONE[7]
Experimental testbed	Realistic wireless environment, realistic hardware and system	UMass DieselNet[9], ORBIT[10]

Table 1. Comparison of approaches and tools in DTN protocol testing and evaluation: from simulation to experiment.

abundant mobility environments to enable the experimenters to repeat their different evaluations. Third, TUNIE is remotely accessible and sharable by the research community, which substantially reduces the capital costs and human effort required to perform experiments. We implement a prototype of TUNIE to demonstrate that it offers a flexible platform for simulating realistic DTN environments and evaluating DTN protocol performance.

The rest of this article is organized as follows. After reviewing the current testing tools and platforms, we provide the design goals of TUNIE and describe the experiment workflow of TUNIE. We then describe the deployment of TUNIE and conduct some preliminary experiments in our implemented TUNIE to show its flexibility as a DTN performance evaluation platform. We conclude the article in the final section.

From Simulation to Experiment

The performance of a DTN may vary significantly, depending on how the mobile nodes move, how densely the nodes are distributed, and how far apart the sender and receiver are. The key factors that determine DTN performance are the routing and forwarding algorithms used, and how well their design assumptions match the actual mobility patterns. Many routing schemes have been proposed. Simulation and, subsequently, experiment testing play an important role in evaluating these DTN protocols. Table 1 categorizes the testing tools into simulation and experiment-based classes, and summarizes the main features of different schemes.

Simulation Testing

Simulation testing for DTN protocol evaluation began with the use of general network simulation software, such as NS-2, NS-3 and OPNET. These simulation tools are designed for general networks, which also include wireless and mobile networks. Therefore, they have been used for DTN protocol evaluation, particularly in the early days. As mobility patterns are important for characterizing DTNs, mobility generators based on simple models are available for NS-2 and NS-3 as part of their toolsets or as specific extensions.

OMNeT++ [8] is a public source simulation platform that has primarily been used for simulating communication networks, and [8] proposes mechanisms for simulating DTN in the OMNeT++ discrete event simulator. These mechanisms allow open systems of wireless mobile nodes to be simulated, where mobility or contact traces are used to drive the simulation. In this approach, the mobility generations, which are separate from the core OMNeT++ protocol simulations, facilitate importing synthetic or real data from external mobility generators, real mobility tracking data, or real contact traces.

While NS-2 and OPNET can offer sound generic open simulation platforms for packet-based communications, and OMNeT++ is embedded with the specific settings to simulate node mobility, generic support for DTN simulation in these platforms is fairly limited. To alleviate this drawback, the ONE simulator [7] contributes an environment for DTN protocol evaluation with embedded internal and external mobility models, enabling different DTN routing schemes and interactive inspection. ONE is an agent-based discrete event simulation engine. At each simulation step, the engine updates a number of modules that implement the main simulation functions, which include the modeling of node movements, inter-node contacts, routing, and message handling. Result collection and analysis are done through visualization, reports, and post-processing tools.

Experimental Testing

Using the above-mentioned simulation tools, however, it is difficult to realize realistic DTN environments in terms of both physical properties and wireless network phenomena. This is because the mobility, channel, and radio characteristics, power consumption, and many other features of wireless mobile networks interact in very complex relationships, and these software simulators simply cannot faithfully reproduce these highly complex and interdependent characteristics. To address these challenges, ORBIT [10] and UMass DieselNet [9] have been designed and built for realistic mobile networking experimentation.

The ORBIT radio grid testbed was developed for scalable and reproducible evaluation of next-generation wireless network protocols by providing a flexible, open-access, multi-user experimental facility. The ORBIT testbed consists of an indoor radio grid emulator for controlled experimentation and an outdoor field trial network for end-user evaluations in real-world settings. This testbed currently consists of

The key factors that determine DTN performance are the routing and forwarding algorithms used, and how well their design assumptions match the actual mobility patterns. Many routing schemes have been proposed. Simulation and subsequently experiment testing play an important role in evaluating these DTN protocols.

400 wireless nodes having 802.11a/b/g wireless cards laid out in a 20 × 20 grid. In this way, DTN protocols can be assessed by the experimenter via an Internet portal, which provides a variety of services to assist the user with setting up network topology, programming radio nodes, executing experimental code, and collecting measurements.

DieselNet consists of computer-equipped buses, battery-powered nomadic nodes, organic WiFi APs, and a municipal WiFi mesh network serving the area surrounding the University of Massachusetts, Amherst campus. More specifically, UMass DieselNet consists of 40 buses, each carrying a small-form desktop computer with 40 GB of storage and a GPS device. Each bus operates a 802.11b radio that scans for other buses 10 times a second and an 802.11b access point that accepts incoming connections. It is a realistic vehicular DTN testbed, and protocols to be evaluated can be implemented on DieselNet as the first step toward real-world deployment. Moreover, testing on DieselNet allows the experimenter to study the effects of certain critical events, such as delays caused by computation, wireless channel interference, and operating system delays, which could not be perfectly modeled in a software simulator.

Trends

Because software-based simulation has difficulties in realizing realistic wireless link and node mobility properties, it is hard for a software simulator to emulate a complicated DTN environment. Recently, emerging real-life DTN testbeds, like DieselNet and ORBIT, offer realistic DTN environments, and they are built to carry out experiments for evaluating DTN related algorithms and protocols. However, setting up an experiment in such a testbed involves huge costs; moreover, the wider research community may not have access to these testbeds. Furthermore, it is very difficult to set up a truly large-scale experiment in these testbeds. It is highly desired that DTN testing platforms are supported by real-world large-scale mobility traces and are flexible to realize. On the other hand, the recently emerging testbed MoViT [11] shows that emulation is an effective and useful approach to evaluate mobile networks. Based on this trend and the above motivations, we propose a large-scale DTN testbed, TUNIE, which exploits virtualization technology and OpenFlow [12] to implement a realistic and reliable DTN environment, while providing remote accessing and sharing for the research community to emulate customized system-level parameters. Thus, TUNIE offers a realistic testing environment and is convenient to use.

TUNIE System Design

TUNIE stands for Tsinghua University Network Innovation Environment. Our design focuses on providing a reputable and controllable emulation testbed for accurate DTN protocol evaluation using the technology of network virtualization. In this section, we first describe the system goals, and then provide the details of the system in terms of controller, node and link design.

TUNIE stands for Tsinghua University Network Innovation Environment. Our design focuses on providing a reputable and controllable emulation testbed for accurate DTN protocol evaluation using the technology of network virtualization.

Design Goals

The primary design of TUNIE is to enable controllable and reliable performance evaluation for DTN protocols. Specifically, TUNIE is designed to achieve the following goals.

High Credibility — To avoid inaccurate evaluations of DTN protocols, which may happen when using simulation tools like NS-2 and ONE, TUNIE should provide an accurate assessment approach that reflects the realistic DTN environment. This environment should include realistic node mobility scenarios, wireless link layer properties, and system-level parameters in network layer implementations. The best choice would be to use a realistic wireless testbed. However, the huge cost in terms of time and money involved makes it unrealistic to implement such a wireless testbed. In TUNIE, we rely on a virtualized testbed to implement realistic DTN environments by controlling the data transmission through regulated wireless links, which exhibit intermittent connectivity with time-varying interference, bit rates, error rates, and transmission delays. At the same time, these realistic experiment environments can easily be configured by remotely accessing the central control center.

System-Level Parameter Realization — New DTN algorithms, protocols, and architectures often require specific customizations of system-level parameters. Therefore, the experiment platform must provide deep programmability of networking functions to implement and emulate system-level parameters, such as operation system (OS) information, network stack settings, distributed protocol states, and interactions on different modules. Consequently, TUNIE should offer sufficient customization to enable the implementation of these system-related parameter details in terms of network stack and transmission links in the deployment of new algorithms and protocols. In the implementation of TUNIE, we use virtual machines supporting a customized OS to run full network stack implementation with new routing protocols and applications, which process the network packets with system-level parameters of processing delay and buffer size, and use a centralized control center to control the link behavior according to different mobility models, and passes through a wild wireless interface as required by the experimenter, which results in real-world link parameters of packet loss rate, transmission delay, and so on.

Remotely Accessible and Repeatable Experiments — In the DTN protocol evaluation, we need to validate the performance of the new algorithms using different mobility scenarios. Therefore, a testbed emulator should provide abundant mobility environments to enable the experimenters to repeat their different evaluations and experiments, in which similar environments can be repeated to obtain similar results. Moreover, it should be remotely accessible to provide the DTN research community with a convenient and practical platform to use. TUNIE offers the experimenter a wide range of choices

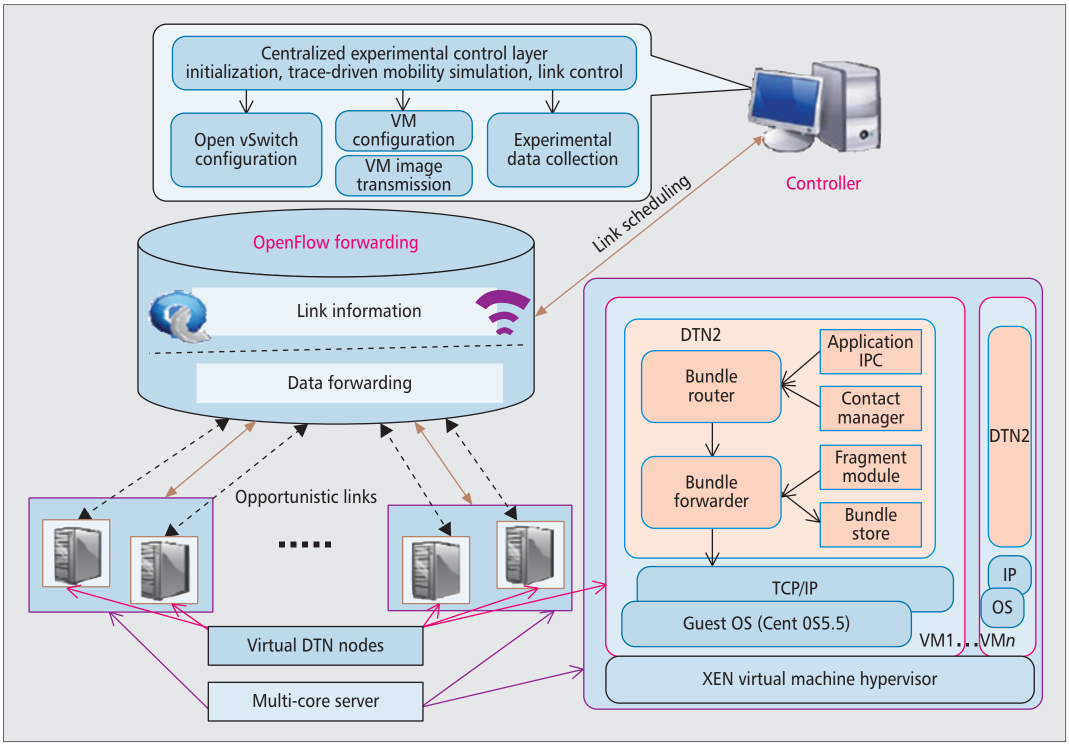


Figure 1. TUNIE architecture: network virtualization based DTN emulation testbed overview.

To enable large-scale experiments and sharable access by many users, TUNIE uses the operation system based network virtualization technology to allow many logical nodes to operate on the same shared physical infrastructure for efficient utilization of the available infrastructure among different experiments.

in terms of mobile environments, and gives the experimenter a high level of control over protocols and software used on the network nodes. In particular, it allows the experimenter to implement a customized system by remote operation and to deal with the issues occurring during the emulation robustly.

Scalability and Shareable Access — Properly evaluating a DTN protocol requires testing its scalability by changing the system size. With the existing testbeds, it is hard to support a large number of nodes, say thousands. On the other hand, as an evaluation platform, a testbed emulator should be a public facility to be shared among many researchers. To enable large-scale experiments and sharable access by many users, TUNIE uses operation system based network virtualization technology to allow many logical nodes to operate on the same shared physical infrastructure for efficient utilization of the available infrastructure among different experiments. The virtual nodes behave like realistic nodes with real-life node mobility and system parameters of network layer implementations, and a unified management system controls the network resource for effective node allocation and usage by all experiments. Due to TUNIE's remote access and sharing capability, the capital costs and human effort required to perform an experiment are substantially reduced, which makes our testbed emulator more economical to run.

System Overview

In order to support large-scale experiments and node customization for credible DTN protocol evaluation, we use the OS and network virtualization technologies to set up virtual DTN emulation platforms. In every DTN node of the TUNIE architecture, an integral of protocol stacks is deployed, which includes the DTN bundle router and bundle forwarder modules, based on the OS visualization. Therefore, the experimenter can easily set up typical routing and forwarding protocols in a DTN node, such as epidemic routing, two-hop forwarding, and spray and wait, or other customized routing and forwarding protocols. To simulate realistic communication links in the DTN paradigm, which relies on intermittent opportunistic connection to transmit packets, we use OpenFlow [12] to control the link events, such as link up and down, according to different inputs of the mobility scenarios. The bandwidths of the transmission links are controlled by the centralized controller, and they vary with time and space. Moreover, we further use the link virtualization technology to control each real-time transmission on the opportunistic links. In this way, we make a virtual wireless interface behave like a real wireless link, including the realistic behaviors of time-varying interference, bit rate, error rate, and transmission delay. This is in contrast to many existing DTN simulators, which either neglect the details of wireless link characteristics and simply assume that any two nodes can communicate with each other when they are in the transmission range of each other, or use the simple time-varying link transmission model to simulate the wireless transmission behaviors as in ONE [7].

The TUNIE architecture is shown in Fig. 1, where two important features can be observed. First, we use OS virtualization to emulate DTN nodes. Specifically, we use a XEN virtual machine hypervisor to run a series of virtual machines as DTN nodes. Each DTN node contains a CentOS supported realistic network stack that includes the upper layer of TCP/IP and core

Data transmissions in TUNIE undergo real-word packet loss rate, transmission delay, and throughput fluctuation. Thus, these important system-level parameters are built in naturally to help obtain accurate and correct performance results for the DTN protocol evaluations.

layer of DTN, which allow users to program each node to customize their own designed algorithms and protocols. Second, OpenFlow is used to control the links of all DTN nodes in order to ensure that they behave like the required real opportunistic links. Specifically, a controller running in a PC transforms the mobility settings given by the experimenter into the link up and down events, and sends these events to the OpenFlow switches, which connect all the virtual nodes through wired and wireless links. The OpenFlow switches control the connectivity between all node pairs to ensure that the transmissions occur in the required opportunistic way, and the wild wireless transmissions ensure they behave as they should in a realistic wireless network. Next, we detail the important components of TUNIE, including node architecture, link structure, and controller.

Node Design

The DTN node in TUNIE is a software-based solution that provides the virtual environment for running a DTN protocol stack and specific applications for the required protocol evaluation settings. As shown in Fig. 1, we use XEN virtualization technology [13] in our system, where multiple concurrent virtual machines running the Linux OS CentOS coexist in a physical machine, and a DTN node is implemented as a virtual machine. In the virtual machine, the core component is the DTN protocol stack, which is running on the TCP/IP protocol stack. In our deployment, we use the open source implementation of DTN bundle protocol, DTN2 [14]. In this DTN stack, Bundle Router and Bundle Forwarder are the two most important components to implement the bundle protocol. Specifically, Bundle Router makes the routing decision based on the application requirements and contact information generated by the contact manager. In the virtual machine, most general router algorithms, including epidemic routing, two-hop relaying, and spray and wait [3], are available to provide the routing functionality. Through the user configuration interface, experimenters can choose their favorite routing components or implement a new one. The physical machine host provides the software interface between the processes in different virtual machines. For a Bundle Forwarder, we use the Click [15] engine to configure the router decision from the Bundle Router into the Click data plane. It then forwards the data into the network interface through the bundle store according to the decision of the fragmentation module. All the packet processing, including the TCP/IP layer packet disassembly and congestion control, will induce delays in packet forwarding. At the same time, the DTN stack with a buffer for message storing, and the buffer size, which is an important system parameter for determining the system performance, are open for user configuration. TUNIE is able to emulate a real DTN system with all these system-level features of the network stack, which are impossible or hard to capture by simulation tools and analytical methods.

The host uses an OpenFlow vSwitch (OVS) [16] to connect virtual nodes, to be discussed later. The OVS can control the bandwidth of each virtual node to share the physical link capacity. After receiving the link bandwidth sent by the controller according to the transmission scheduling, the host can dynamically change the bandwidth. This control mechanism is enabled by the XEN network interface virtualization scheme, where the host can view and control the virtual interface in each virtual node.

Link Design

The virtual link is a key component of TUNIE. We use the link virtualization technology to enable the emulation of wireless opportunistic links. We first set up a virtual interface, which is a tap device, for each virtual node. We then use the software bridge of the OVS in the XEN virtualization environment to connect all the tap devices with the WiFi physical interface in the host node, which itself is connected to the OpenFlow access point (AP). The OVS and OpenFlow AP together let any two virtual nodes set up a link at any time, and can also block any connection at any time. The link control is achieved by the flow management in OpenFlow, which is further controlled by the centralized controller. The link view of TUNIE is illustrated in Fig. 2, where it can be seen that the OpenFlow AP connects the physical multi-core servers through the WiFi interfaces, and this in turn allows virtual nodes to connect with each other through the wireless links. With the OpenFlow link controlling function, all the links among node pairs are controlled by the system; therefore, it is easy for experimenters to simulate different mobility scenarios. On the other hand, the controller controls all the transmissions over virtual links either across multi-servers or insider one physical server, illustrated by virtual link a and b, respectively, in Fig. 2, as they pass the wild WiFi interfaces. Consequently, wireless characteristics, such as radiation patterns of the antennas, are taken into account in the emulation testbed. In this way, data transmissions in TUNIE undergo real-word packet loss rate, transmission delay, and throughput fluctuation. Thus, these important system-level parameters are built in naturally to help obtain accurate and correct performance results for the DTN protocol evaluations.

Controller Design

The controller is a centralized network link control component in the OpenFlow network, which is running on the network OS and has an overview of the network. The controller can inject specific flow rules into the OpenFlow switch according to the behaviors of the network, which depend on the routing algorithm, load balance, and so on. Our TUNIE controller is built on OpenFlow NOX, and is implemented in Java and Python languages. Specifically, we develop a centralized experiment control layer in NOX. The input parameters of this control layer include the number of nodes, mobility model, and emulated applications. Based on the given input, the control layer provides the functions of VM Image transmission and configuration to set up virtual DTN nodes, open vSwitch configuration to control the links, and experimental controlling and data collection, as shown in Fig. 1. To control the experiment, for instance, it first

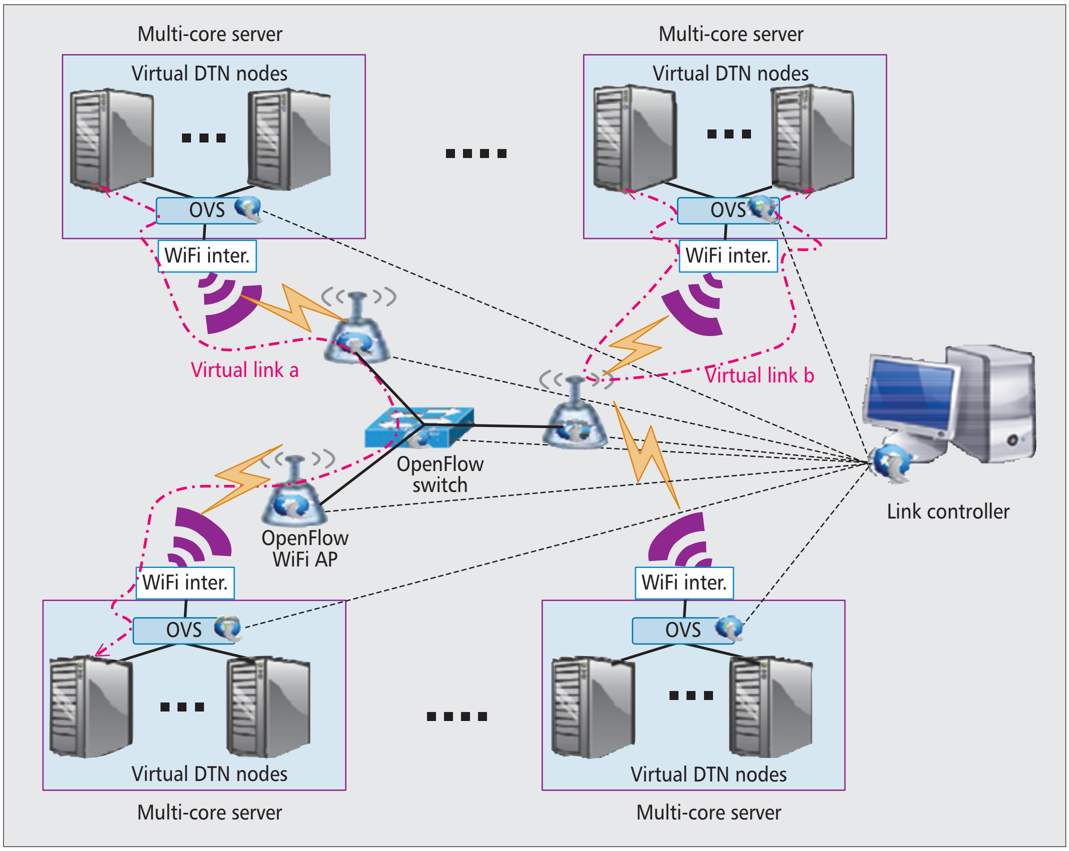


Figure 2. Link design and virtual links in TUNIE.

Related to the routing protocols, users can choose epidemic routing, two-hop relaying, and spray and wait, which are available in the system for configuring the DTN nodes. Users can also implement new routing and resource allocation schemes to evaluate their own protocols.

computes the opportunistic links among all nodes based on the given mobility model to obtain the specific link up and down events. Second, it calculates the data transmission rate of each opportunistic link according to the positions of the communicating nodes. After obtaining the link events and transmission rates, it computes the flow rules at different times, and then sends them to the OpenFlow switch at the appropriate times, as well as deleting certain rules after their link times are done. With this mechanism, we can use the OpenFlow switch to control the links of different nodes to emulate an opportunistic DTN.

TUNIE Experiment Working Flow

The experiment working flow of TUNIE is depicted in Fig. 3. When setting up an experiment in TUNIE, the user needs to configure the controller first by setting the overall network parameters, including the number of nodes, mobility model, and so on. TUNIE will generate the virtual nodes by communicating with the required multi-core servers. The user then needs to log in to the virtual nodes to configure and prototype the algorithms and protocols. The TUNIE controller will translate the experiment settings into link control events and, through the OpenFlow flow rules, controls the links among virtual nodes to ensure that they behave as the required opportunistic links. We now illustrate the experiment working flow by introducing our mobility scenarios, and routing and application settings implemented in TUNIE.

Mobility Scenarios

Mobility is the most important consideration for DTN performance evaluation. Usually, different performance evaluations of a protocol require different mobility models. For example, to evaluate the scalability properties, we may adopt a random mobility model, with which the scale of the network is easy to change. On the other hand, to evaluate the transmission efficiency, we need real traces covering different mobility scenarios in order to get accurate results. To satisfy these different requirements, in TUNIE we integrate different mobility scenarios in the controller, including random mobility models, map-based mobility, and real-world human and vehicular mobility traces, as indicated in Fig. 3.

In terms of random mobility, we cover a broad set of random mobility models by including random waypoint, random walk, and random direction walk. For map-based mobility, we use a map model in the ONE simulator [7]. For realistic human mobility, we integrate four human mobility traces, *Infocom05*, *Infocom06*, *Reality*, and *Cambridge*, in the system. Among these four human mobility traces, Reality was collected from the MIT Reality Mining Project, and the other three were gathered by the Haggle Project. We also integrate two vehicular mobility traces, *Shanghai* and *Beijing*. The Shanghai trace was collected by involving 2019 operational taxis in Shanghai over the whole month of February 2007 without any interruptions. The Beijing

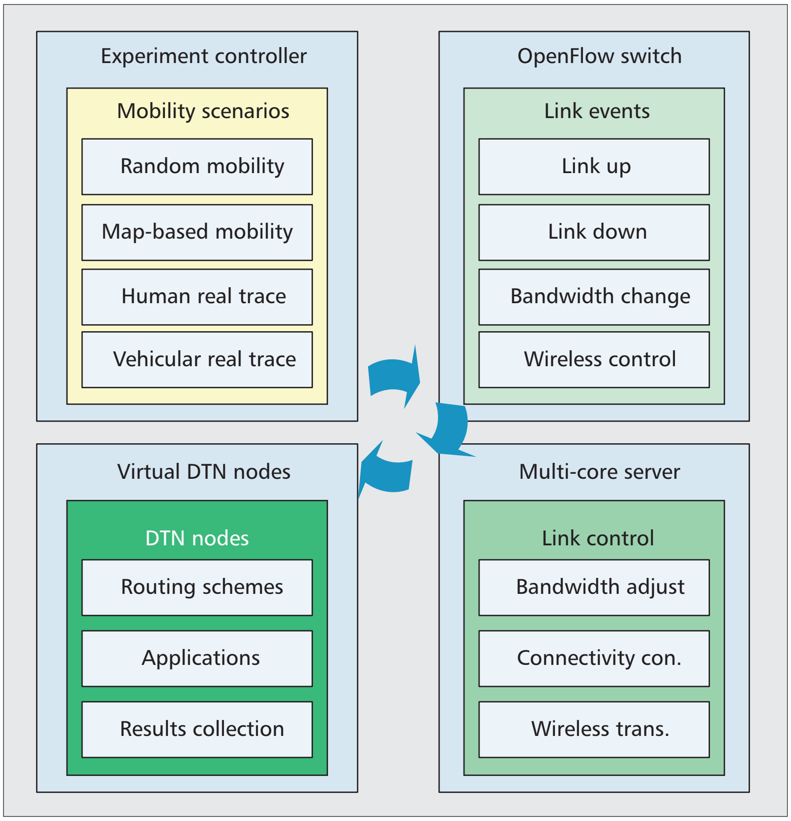


Figure 3. System functions and experiment workflow of TUNIE.

trace includes the mobility traces of 27,000 participating Beijing taxis collected during the entire month of May in 2010, which is the largest vehicular data trace available.

Routing and Applications

Related to the routing protocols, users can choose epidemic routing, two-hop relaying, and spray and wait, which are available in the system for configuring the DTN nodes. Users can also implement new routing and resource allocation schemes to evaluate their own protocols. All these functions are openly accessible in the DTN Bundle module of the DTN stack. In terms of applications, similar to the ONE simulator, TUNIE provides two ways to generate application packets in the virtual nodes: packet generators and external traffic event files. The packet generator in TUNIE creates packets for the selected source and destination with the given packet size and deadline, which are set by the user. A separate tool to generate packet event files is also included in TUNIE. In this way, users can generate messages very conveniently. For both ways of generating application packets, the application is implemented in the module of the Application IPC, which communicates with the router (i.e., injects bundles into the Bundle Router) via an inter-process communication channel in the experiment control layer.

Implementation and Deployment

In implementing the TUNIE design on hardware devices, we use 40 high-performance servers with Intel Xeon X5550 2.6 GB four core CPU, 16 GB memory, and WiFi interface as the hosts to provide the virtual DTN nodes. For the OpenFlow switches, we use 8 WiFi APs of a Broadcom chip and 3 Pronto OpenFlow-enabled switches as the OpenFlow devices, where the WiFi APs provide the wireless connections of the servers through their WiFi interfaces. The system-level parameters and characteristics are summarized in Table 2, which provides detailed information, in terms of nodes, links, and DTN protocol stack, on our implemented testbed emulator. This virtualized programmable DTN emulation testbed contains the resources of software virtualization component and OpenFlow component. With our unified user configuration interface in the controller and host, experimenters have access to all the resources in the platform, and can build their own experiments to test the designed DTN algorithms and protocols. Moreover, the same implemented code of these evaluated targets can directly run on a real-life DTN system, and no translation is required from the testbed emulator to the real DTN environment.

Based on the virtualized environment, we install the virtual DTN nodes with the DTN2 protocol stack on the TCP/IP network stack, and we also install a network connecting with all the physical multi-core servers to support the platform management system. We design the system with the GUI interface by web service for users to apply the DTN node resources, and to use them by configuring and customizing the network. All these operations and configurations over the virtual environment are carried out by web-based remote access, and experimenters are able to build their own experiments concurrently in their individual and separated virtual networks. Although building TUNIE involves the cost of the required hardware servers and switches, it is significantly less than deploying a similar large-scale dedicated hardware testbed. Moreover, such a dedicated testbed can only run a single experiment, but our TUNIE testbed emulator enables multiple different experiments to run concurrently, and its hardware resources can be remotely accessed, controlled, and shared by many researchers. Additionally, unlike a dedicated DTN testbed, the invested hardware resources are not restricted to DTN experiments, and TUNIE can conveniently be used for other network experiments and application. Clearly, our TUNIE testbed emulator provides an economical, repeatable, and reliable platform for DTN protocol testing and evaluation, and it will be open to the wider DTN research community.

We carried out some basic experiments to investigate the capability of the system. The well-known ODE model is widely used to model the message propagation in DTN network [5]. Here, we verify this model in our testbed emulator to investigate the efficiency and accuracy of TUNIE as a DTN performance evaluation platform. We evaluated the accuracy of the ODE model by comparing the theoretical results obtained based on the model given in [5] with the experimental results, which were obtained by simulating the message dissemination under the epidemic routing with the random walk mobility model in TUNIE. The system settings and performance metrics obtained from the experiment on

	Category	Parameters
Node virtualization	Physical server	Intel X5550 2.6G 4 core CPU, 16 GB memory, 800 GB hard disc, and Cisco 350 WiFi adaptor.
	Virtual node	1 virtual CPU, 256 MB memory, 4 GB hard disc, virtual tap network interface, CentOS.
	Virtualization efficiency	32 virtual nodes/physical server, full virtualization, CPU utilization efficiency > 80 %.
Link virtualization	Wired port and link	Tap rate = 10,000 ± 100kb/s, Open vSwitch port rate control: max = 20 Mb/s, min = 0.5 Mb/s.
	Wireless links	Frequency: 2.4–2.4897GHz; Data Rates: 1, 2, 5.5, and 11 Mb/s; transmit power: 0-20dBm.
DTN protocol stack	Routing protocols	Static, flooding, two-hop relaying, spray and wait, delay-tolerant link state routing.
	Transmission protocols	TCP, UDP, Ethernet, BlueTooth, serial.
	Storage size setting	Bundle storage: 2 MB–10 GB; message storage: 16 KB–2 MB.

Table 2. System-level parameters and characteristics of our deployed TUNIE testbed emulator.

TUNIE in terms of one-hop transmission and network-level performance are summarized in Table 3.

We simulated the network with different network sizes of *N* from 100 to 1000, where 20 percent of the nodes in the network were randomly chosen to be infected, and the system was simulated 100 times. From the average deviation between the theoretical and experimental results of the average infected node ratio in Table 3, we inferred that the number of message-infected nodes computed from the ODE model agreed with the average values obtained by the experiment. On the other hand, the deviations, which were influenced by TUNIE's system-level parameters of network stack and wireless links, indicated that the deviations happened in realistic DTN environments with network stack interactions and wild wireless transmissions. These aspects could not be obtained by the analytical method, which only captured the average system performance. These results clearly demonstrate the credibility and accuracy of TUNIE as an emulation testbed to evaluate DTN performance.

Category	Metrics	Values
One-hop performance	Wireless transmission rate	5.5 Mb/s
	Average wireless link throughput	2.4 Mb/s
	Throughput standard deviation	1.3 Mb/s
	Link packet loss rate	9.5%
Network performance	The number of nodes (*N*)	100~1000
	System simulated time (*T*)	5×10^3 s
	Average message delivery ratio	91.2%
	Average message transmission delay	1.7×10^3 s
	Average deviation of infected ratio	99% ± 8%

Table 3. System-level settings and performance of the experiment of ODE model validation in TUNIE.

Conclusions

In this article, we have reviewed the evolution of DTN protocol testing and evaluation. Based on a survey of the existing simulation tools and experimental testbeds, particularly their advantages and drawbacks, we have designed and implemented TUNIE, a network virtualization based DTN testbed emulator. TUNIE integrates XEN virtualization and OpenFlow technologies. Specifically, XEN enables setting up large-scale DTN networks, and OpenFlow is used to emulate the DTN opportunistic links. TUNIE opens up a wide range of choices to users, in terms of mobility scenarios, routing protocols, and applications, and also allows users to design their own experimental environments conveniently. Our initial implementation validated that TUNIE is an efficient platform to evaluate DTN performance, which needs to further integrate other abundant models about the network properties (i.e., link models from MoViT[11] and device characteristics (i.e., energy consumption) to simulate some specific DTN environments. TUNIE's web-based interface for remotely accessing, controlling, and sharing ensures that TUNIE will be open to the wider DTN research community.

References

[1] K. Fall, "A Delay-Tolerant Network Architecture for Challenged Internets," *Proc. ACM SIGCOMM 2003 Conf. Applications, Technologies, Architectures, and Protocols for Computer Commun.*, Karlsruhe, Germany, Aug. 25–29, 2003, pp. 27–34.

[2] K. Fall and S. Farrell, "DTN: An Architectural Retrospective," *IEEE JSAC*, vol. 26, no. 5, June 2008, pp. 828–36.

[3] Z. Zhang, "Routing in Intermittently Connected Mobile Ad Hoc Networks and Delay Tolerant Networks: Overview and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 8, no. 1, 2006, pp. 24–37.

[4] Y. Li *et al*., "Evaluating the Impact of Social Selfishness on the Epidemic Routing in Delay Tolerant Networks," *IEEE Commun. Lett.*, vol. 14, no. 11, Nov. 2010, pp. 1026–28.

[5] X. Zhang *et al.*, "Performance Modeling of Epidemic Routing," *Computer Networks*, vol. 51, no. 10, Oct. 2007, pp. 2867–91.

[6] S. Jain, K. Fall, and R. Patra, "Routing in a Delay Tolerant Network," *Proc. ACM SIGCOMM 2004 Conf. Applications, Technologies, Architectures, and Protocols for Computer Commun.,* Kyoto, Japan, Aug. 30–Sept. 3, 2004, pp. 145–58.

[7] A. Keränen, J. Ott, and T. Kärkkäinen, "The ONE Simulator for DTN Protocol Evaluation," *Proc. 2nd Int'l. Conf. Simulation Tools and Techniques for Commun., Networks and Sys.*, Rome, Italy, Mar. 2–6, 2009, pp. 50–55.

[8] O. R. Helgason and K. V. Jónsson, "Opportunistic Networking in OMNeT++," *Proc. ACM 1st Int. Conf. Simulation Tools and Techniques for Commun., Networks and Sys.*, Marseille, France, Mar. 3–7, 2008, pp. 76–82.

[9] A. Balasubramanian, B. N. Levine, and A. Venkataramani, "DTN Routing as a Resource Allocation Problem," *Proc. ACM SIGCOMM 2007 Conf. Applications, Technologies, Architectures, and Protocols for Computer Commun.*, Kyoto, Japan, Aug. 27–31, 2007, pp. 373–84.

[10] W. Ivancic *et al.*, "Experience with Delay-Tolerant Networking from ORBIT," *Proc. IEEE 4th Conf. Advanced Satellite Mobile Sys.*, Bologna, Italy, Aug. 26–28, 2008, pp. 173–78.

[11] E. Giordano *et al.*, "MoViT: the Mobile Network Virtualized Testbed," *Proc. ACM VANET '12*, Lake District, U.K., June 25, 2012, pp. 3–12.

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

[13] P. Barham et al., "Xen and the Art of Virtualization," *ACM SIGOPS Op. Sys. Rev.*, vol. 37, no. 5, Dec. 2003, pp. 164–77.

[14] S. Burleigh, "Interplanetary Overlay Network: An Implementation of the DTN Bundle Protocol," *Proc. IEEE Consumer Commun. and Networking Conf.*, Las Vegas, NV, Jan. 11–13, 2007, pp. 222–26.

[15] E. Kohler et al., "The Click Modular Router," *ACM Trans. Computer Sys.*, vol. 18, no. 3, Aug. 2000, pp. 263–97.

[16] B. Pfaff et al., "Extending Networking into the Virtualization Layer," *Proc. 8th ACM Workshop on Hot Topics in Networks*, New York, NY, Oct. 22–23, 2009, pp. 1–6.

BIOGRAPHIES

YONG LI [M'09]received his B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and his Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. During July to August 2012 and 2013, he was a visiting research associate with Telekom Innovation Laboratories and Hong Kong University of Science and Technology, respectively. During December 2013 to March 2014, he was a visiting scientist with the University of Miami, Florida. He is currently a faculty member of the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of networking and communications, including mobile opportunistic networks, device-to-device communication, software-defined networks, network virtualization, and future Internet.

PAN HUI received his Ph.D degree from the Computer Laboratory, University of Cambridge, and earned his M.Phil. and B.Eng. from the Department of Electrical and Electronic Engineering, University of Hong Kong. He is currently a faculty member of the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology, where he directs the System and Media Lab. He also serves as a Distinguished Scientist of Telekom Innovation Laboratories (T-labs) Germany and an adjunct professor of social computing and networking at Aalto University, Finland. Before returning to Hong Kong, he spent several years in T-labs and Intel Research Cambridge. He has published more than 100 research papers, and has several granted and pending European patents. He has founded and chaired several IEEE/ACM conferences/workshops, and served on the Technical Program Committees of numerous international conferences and workshops including IEEE INFOCOM, SECON, MASS, GLOBECOM, WCNC, and ITC.

DEPENG JIN received his B.S. and Ph.D. degrees from Tsinghua University in 1995 and 1999, respectively, both in electronics engineering. He is an associate professor at Tsinghua University and vice chair of Department of Electronic Engineering. He was awarded the National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design, and future Internet architecture.

SHENG CHEN [M'90, SM'97, F'08] obtained his B.Eng. degree from the East China Petroleum Institute, Dongying, China, in January 1982, and his Ph.D. degree from City University, London, United Kingdom, in September 1986, both in control engineering. In 2005, he was awarded a D.Sc. from the University of Southampton, United Kingdom. From 1986 to 1999, he held research and academic appointments at the Universities of Sheffield, Edinburgh, and Portsmouth, all in the United Kingdom. Since 1999, he has been with the Department of Electronics and Computer Science, University of Southampton, where he currently holds the post of professor in intelligent systems and signal processing. He is a distinguished adjunct professor at King Abdulaziz University, Jeddah, Saudi Arabia. He is a Chartered Engineer (CEng) and a Fellow of IET (FIET). His recent research interests include adaptive signal processing, wireless communications, modeling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, evolutionary computation methods, and optimization. He has published over 470 research papers. He is an ISI highly cited researcher in the engineering category (March 2004).

On the Potential of Bluetooth Low Energy Technology for Vehicular Applications

Jiun-Ren Lin, Timothy Talty, and Ozan K. Tonguz

Jiun-Ren Lin and Ozan K. Tonguz are with Carnegie Mellon University.

Timothy Talty is with General Motors LLC.

ABSTRACT

With the increasing number of sensors in modern vehicles, using an intra-vehicular wireless sensor network (IVWSN) is a possible solution for the automotive industry for addressing the potential issues that arise from additional wiring harness. Such a solution could help car manufacturers develop vehicles that have better fuel economy and performance, in addition to supporting new applications. However, which wireless technology should be used for maximizing the benefits of IVWSNs is still an open issue. In this article we propose to use a new wireless technology known as Bluetooth Low Energy (BLE) and outline a new architecture for IVWSN. Based on a comprehensive study that encompasses an example application, it is shown that BLE is an excellent option that can be used in IVWSNs for certain applications mainly due to its good performance and low-power, low-complexity, and low-cost attributes.

INTRODUCTION

Modern production vehicles are highly computerized, and the major functionalities of a vehicle are controlled by several electronic control units (ECUs) inside the vehicle. ECUs need to gather information about the vehicle from the sensors in order to maintain all the required vehicular operations. Currently, most of the sensors inside vehicles are connected by physical wires, so each sensor sends out its data via the wires toward its destination ECU. However, because vehicles are becoming more complex, and the number of applications and gadgets in vehicles keeps increasing, the large number of wires needed for the connection of sensors poses several significant challenges, the first one being the extra weight of the wires. If the extra weight can be eliminated, the weight of vehicles can be reduced, and thus they can have better fuel economy and performance. Furthermore, the wired connection limits the possible sensor locations and hence the range of applications. The wires themselves are costly, and the cost of installing wires in vehicles can be high for car manufacturers. When a vehicle gets older, some wires may deteriorate and cause severe problems, and to replace wires inside a vehicle would be either impossible or very expensive. In order to address these issues, wireless technology was recently proposed for communications between sensors and ECUs. The wireless sensors and the ECUs form a new architecture, which is often referred to as an intra-vehicular wireless sensor network (IVWSN) [1].

Because of the potential benefits of IVWSNs, car manufacturers might gradually introduce wireless sensors into vehicles in the near future. Such a gradual scheme could start from several possible types of sensors: those that are not safety-critical, those in hard-to-reach locations, or those that are the easiest to be replaced with wireless sensors. Furthermore, for car manufacturers the additional cost of the wireless hardware is the major barrier for the deployment of IVWSNs. In order to massively deploy wireless sensors in vehicles, the unit price of a sensor with a wireless transceiver should not be much higher than an ordinary sensor. The cost of a wireless sensor highly depends on the chosen wireless technology and the complexity of the system. Consequently, a good starting point is to identify and evaluate a viable wireless technology to support the aforementioned types of sensors. These types of sensors/applications usually have the following requirements and properties.

Requirements

- **Low cost:** Lower complexity implies lower cost. Besides, if the system can adopt an existing wireless technology with minimum modifications, the cost can be further reduced.
- **Low power consumption:** For most wireless sensors, their power is supplied by a battery. Therefore, the power consumption for wireless communications has to be low enough to support a reasonable battery life.
- **Short delay:** For some of the applications, having a short delay (i.e. a few milliseconds) is desirable since the system can be highly dynamic or require prompt response.

0163-6804/15/$25.00 © 2015 IEEE

Bluetooth SIG announced the Bluetooth specification version 4.0 in June 2010. It introduced the new Low Energy (LE) Core Configuration, which is also called Bluetooth Low Energy (BLE) in order to distinguish it from the traditional Basic Rate (BR) and Enhanced Data Rate (EDR) Core Configurations.

- **High reliability:** The system has to provide guaranteed data transmissions.

Properties

- **Low data throughput:** The sensor data are usually very short, that is, only a few bytes.
- **Low duty cycle:** Most of the applications have a low duty cycle, for example, less than five percent.
- **Various priorities:** Depending on the application, different packets are assigned with different priorities. For example, the packets from a safety-critical system generally have a higher priority than the packets from the air-conditioning system.

Moreover, while IVWSNs can be considered as a type of wireless sensor network, IVWSNs have a number of unique characteristics; hence, a specific protocol stack and system design would be required in order to achieve optimal performance. For instance, the sensors in IVWSNs are mostly fixed or can only move within a small area, while classical wireless sensor networks often have a dynamic topology [2]. This implies that node mobility and routing configuration is less of a problem in IVWSNs. However, metal parts, especially in the engine compartment, act as obstacles and create a challenging and unique environment for wireless communications, especially compared to open space environments, as assumed in most of the classical wireless sensor networks. Due to the special physical environment, it is essential to evaluate the wireless technologies for IVWSNs in a bottom-up manner, starting from the physical (PHY) layer.

One of the wireless technologies that could be used for IVWSNs is ZigBee/IEEE 802.15.4. Specifically, it has been shown before that the ZigBee PHY layer is suitable for IVWSNs [3]. However, the investigation in [1] has shown that the MAC protocol of the ZigBee standard may not be suitable for some sensors/applications, and that could imply a customized protocol stack. In fact, since the sensors and applications in a vehicle are heterogeneous (i.e. with different requirements in terms of delay, throughput, duty cycle, and power consumption), it might be necessary to use more than a single wireless technology to fulfill all the requirements of different applications. The future IVWSN might therefore be a hybrid network with multiple wireless technologies coexisting for different groups of sensors. These factors necessitate further research into different options for wireless technologies.

Another possible wireless technology for IVWSN is ultra-wideband (UWB) communications. UWB was introduced in the IEEE 802.15.4a-2007 and IEEE 802.15.4-2011 standards as one PHY option, and other variations of UWB have been studied extensively for the applications within a vehicle [4]. However, while UWB could provide a very large data throughput (several to hundreds of Mb/s) and better resilience to multi-path fading, the cost of UWB technology is still higher than some existing low-power wireless technologies such as ZigBee. Furthermore, for automotive applications, the frequency spectrum of use has to follow the regulations worldwide, and this is one of the major reasons to use a low-power wireless technology that operates in the 2.4 GHz ISM band, which is available worldwide.

In this article we propose to use the Bluetooth Low Energy (BLE) technology [5] as an excellent choice for the IVWSN architecture (the Bluetooth™ word mark and logos are registered trademarks owned by Bluetooth SIG, Inc.). The properties and the performance of BLE will be evaluated specifically for IVWSN applications. With our comprehensive evaluation and discussion, we show that BLE could provide a powerful hardware platform and PHY layer for IVWSN and enable car manufacturers to design and implement IVWSNs with low cost and high efficiency.

The rest of the article is organized as follows. The next section gives an overview of BLE technology. Following that we describe the IVWSN based on BLE and present a detailed comparison between BLE and ZigBee. Then we provide detailed information on the system design, configuration, and the methodology used for an example application: a BLE-based passive keyless entry system. We then discuss the major issues related to the proposed system and applications. Concluding remarks are given in the final section.

Overview of Bluetooth Low Energy

The Bluetooth Special Interest Group (Bluetooth SIG) announced the Bluetooth specification version 4.0 in June 2010. It introduced the new Low Energy (LE) Core Configuration, which is also called Bluetooth Low Energy (BLE) in order to distinguish it from the traditional Basic Rate (BR) and Enhanced Data Rate (EDR) Core Configurations [5]. BLE is designed for applications that have low duty cycle and require low power consumption and low cost. Figure 1a shows the protocol stack of BLE. Note that the Bluetooth core system consists of a host and one or more controllers. A Bluetooth device could have both BR/EDR and LE controllers or only either one.

BLE operates in the unlicensed 2.4 GHz ISM band, and it employs adaptive frequency hopping scheme to combat interference and fading. It uses 40 channels with center frequencies 2402 MHz to 2480 MHz, and each channel is separated by 2 MHz. Among the 40 channels, three are advertising channels and the remaining 37 channels are data channels. BLE uses binary Gaussian Frequency Shift Keying (GFSK) as the modulation scheme, and the symbol rate and bit rate are both 1 Mb/s. The transmitting power of a BLE device is between –20 dBm and 10 dBm.

BLE has two different logical communication groups: one is piconet and the other is broadcast group. In a piconet, there is one master device and multiple slave devices. The maximum number of slaves in a piconet is not defined in the Bluetooth standard, but it is limited by the capabilities of the master device. All communication within a piconet is between the master and slave devices. There is no direct communication between the slave devices in a piconet. In other words, a piconet has a star topology. Before joining a piconet, a slave device can try to join a piconet by broadcasting advertisements on the advertising channels. The master device scans the advertising channels and decides if it wants to establish a connection with the advertising slave device. If the master device allows the advertising slave device to join the piconet, it will initiate the connection to the slave device. After the connection is established, the slave device is synchronized to the timing and frequencies of the physical channel specified by the master device. Note that in a piconet, each slave device uses a different physical channel (i.e. a different frequency hopping sequence) to communicate with the master device.

On the other hand, a broadcast group consists of one advertiser and multiple scanners within the communication range of the advertiser. An advertiser broadcasts advertisements, and scanners scan the three advertising channels and receive the advertisements. There is no continuous connection between the advertiser and the scanners. In other words, while the master and slave devices are doing one-to-one connection-oriented communications in a piconet, the advertiser and scanners are doing one-to-many connectionless communications in a broadcast group.

In a piconet, after the connection is established, there are periodic connection events between the master and each slave device. In a connection event, the master transmits packets to a slave and the slave can respond with a packet depending on the context. Therefore, the master controls the access to the channel in a piconet. Each connection event corresponds to a PHY hop channel. Consecutive connection events correspond to different PHY hop channels. The period of the connection events is defined by the upper layers.

In a BLE host, the generic access profile (GAP) layer (Fig. 1a) controls the device's communication modes and procedures. Depending on the purpose of an application, the GAP layer operates in one of the following four roles: broadcaster (advertiser), observer (scanner), peripheral (slave), and central (master). In addition, a BLE device that operates in the peripheral or central role can also operate in the broadcaster or observer role. The application layer can control the operation role of the device by calling GAP API functions.

For packets in a connection event, each link layer packet uses a 24-bit cyclic redundancy error check (CRC) to cover the payload. If the CRC verification fails at the receiver, the packet will not be acknowledged and the sender will retransmit the packet. On the other hand, there is no acknowledgment or CRC field for the advertisement packets (broadcast packets). Each advertisement is transmitted several times to increase the probability that the scanner can successfully receive at least one of the copies. The length of a regular BLE packet is between 10 bytes and 47 bytes (Fig. 1b); the length of a BLE advertisement packet is between 8 bytes and 39 bytes.

The latest Bluetooth specification to date is version 4.1, which was announced in December 2013 [6]. The major enhancement of the LE portion in Bluetooth version 4.1 is the additional link layer topology support. Bluetooth specification version 4.0 assumes that an LE slave device is only able to join one piconet at a time, but in Bluetooth version 4.1 an LE slave device can also act as a master or slave device of another piconet. Therefore, a scatternet topology is allowed in the new specification.

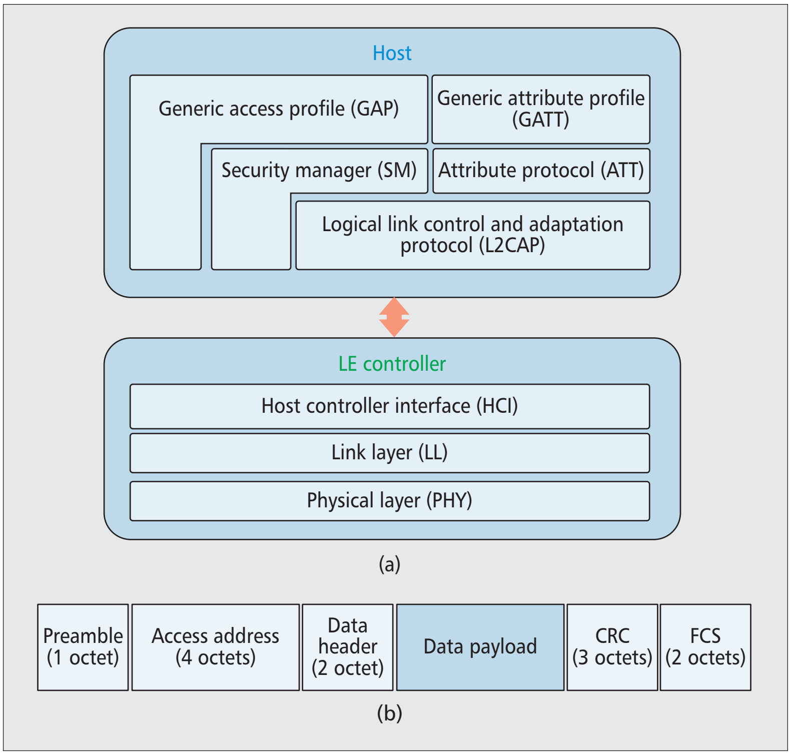


Figure 1. The protocol stack and frame format of Bluetooth Low Energy: a) protocol stack; b) frame format.

IVWSNs Based on Bluetooth Low Energy

According to the existing literature, the intra-vehicular wireless channels have several properties [7]:

- The 90 percent coherence bandwidth at 2.4GHz is around a few MHz, which is at least as large as some indoor channels.
- The coherence time of the intra-car channels ranges from 2.5 seconds to a few hundred seconds, depending on different driving scenarios.
- Huge path losses (e.g. > 80 dB) can be observed when the transmitter and the receiver are in different compartments.

Along with the aforementioned requirements of sensors/applications inside vehicles, the candidate wireless technologies have to be low-power, low-cost, and occupy less than a few MHz of bandwidth. As mentioned previously, BLE is designed for applications that have low duty cycle and require low power consumption and low cost, and the channel bandwidth of BLE is 2 MHz, which is narrower than the coherence bandwidth inside the vehicle. These imply that BLE could also be suitable for IVWSNs as well.

Since BLE was not originally designed for vehicular applications, we conducted a series of

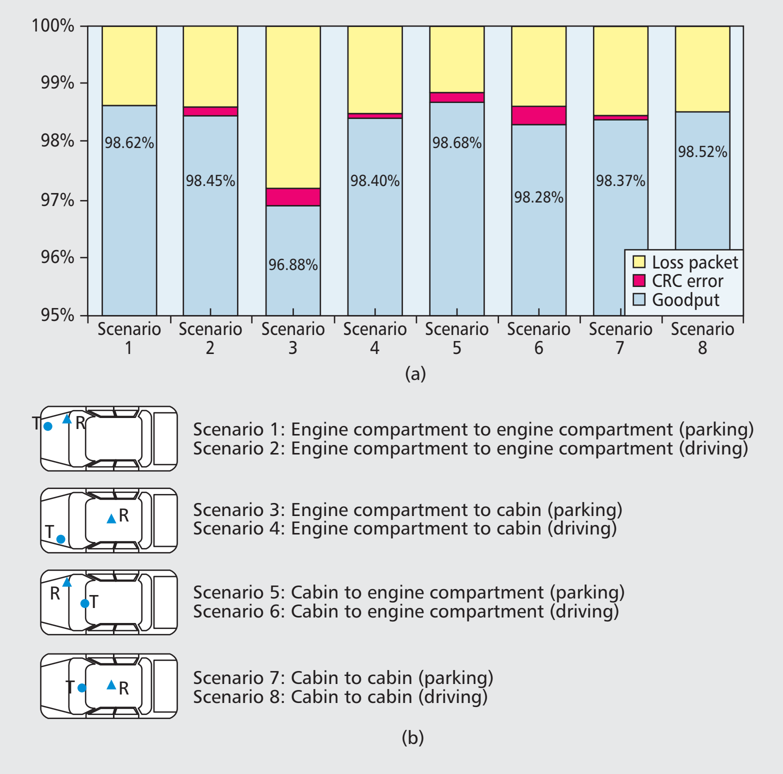


Figure 2. The packet goodput of BLE in eight different intra-vehicular scenarios: a) packet goodput; b) intra-vehicular scenarios.

experiments in order to evaluate the actual performance of the PHY layer of BLE in an intra-vehicular environment. As part of the results reported in [8], it was shown that BLE can provide reasonably good packet goodput in the eight different intra-vehicular scenarios (Fig. 2a). Figure 2b illustrates the positions of the BLE transmitter (denoted by **T**, the transmission power was 0 dBm) and receiver (denoted by **R**) in each scenario.

Performance Parameters

In addition to packet goodput, other important considerations for the wireless technology for IVWSNs are the throughput and delay performance. Since an IVWSN is mainly designed for sensor data communications, a large data throughput might not be necessary. However, if the technology provides more PHY layer throughput, the network will have a larger capacity to accommodate more sensors and data. This is important as car manufacturers are adding more and more features and sensors to modern vehicles. The data communications of BLE is performed in the predefined 37 data channels. The system can support multiple concurrent data communications if each master-slave pair applies an orthogonal hopping sequence. Theoretically, the maximum PHY layer throughput of the entire system could be up to 37 Mb/s if all of the hopping sequences and traffic are carefully arranged. Note that the actual data throughput would depend on the payload size of the sensor packets and the MAC scheduler design.

Compared to data throughput, the delay performance plays a more important role for many automotive applications. Delay is normally measured from the moment that a sensor sends out a data packet to the time the destination ECU receives the packet. For IVWSNs, some sensor data have to arrive at the destination ECUs within a few milliseconds to maintain normal operation of the vehicle. The overall delay consists of three parts: transmission delay, queueing delay, and propagation delay. The transmission delay directly depends on the link data rate and packet size. Since sensor packets are usually fairly short, the major factor that affects the delay performance is the queueing delay. For instance, if the packet size is 20 bytes (i.e. with 8 bytes payload), the transmission delay is 0.16 ms since the data rate of BLE is 1 Mb/s. The propagation delay is about a few nanoseconds, depending on the dimension of a vehicle, and hence could be ignored in most cases. In BLE, slave devices can only send a packet to the master device during the connection events after receiving a packet from the master. The queueing delay is the delay incurred while waiting for the connection event in order to send the sensor packet. Therefore, the connection event has to be carefully scheduled according to the sensor reading time in order to minimize the queueing delay.

Comparison between Bluetooth Low Energy and ZigBee

Several existing works on IVWSNs focused on ZigBee wireless technology [1, 3, 9, 10]. ZigBee is designed for RF applications that require low power consumption, low complexity, and low data rate [11]. The PHY and MAC layers of ZigBee are based on the IEEE 802.15.4-2003 standard. Similar to ZigBee, Bluetooth, another Personal Area Network (PAN) technology, also operates in the 2.4 GHz unlicensed ISM band. According to the conclusions in [12], Bluetooth Basic Rate (BR) and ZigBee are both suitable for low data rate applications with limited battery power. However, Bluetooth BR still consumes more power and has higher complexity than ZigBee does. The main motivation for using BLE is therefore to provide a better solution for low-power and low-cost applications.

Table 1 is a detailed comparison between BLE and ZigBee in terms of several important characteristics. Observe that they have many similarities: both of them operate in the 2.4 GHz ISM band, and the bandwidth of each channel is the same (i.e. 2 MHz). However, since they use different modulation and spreading schemes, their maximum data rates are different: BLE can achieve up to 1 Mb/s data rate, which is higher than ZigBee's 250 kb/s. Another important advantage of BLE is the lower hardware cost. Both BLE and ZigBee are designed to be low-cost technologies, but the unit price of a BLE compliant chip is currently less than a ZigBee compliant chip. A possible reason might be that there are more phones and laptops supporting BLE as part of the Bluetooth 4.0 standard, so it has a larger market than ZigBee does. It also implies that there will be more and more consumer devices that will support BLE in the near future, and it can enable new features on vehicles with lower cost.

	Bluetooth Low Energy	ZigBee
IEEE standard	None	802.15.4-2003
Frequency band	2.4 GHz	868/915 MHz; 2.4 GHz
Max data rate	1 Mb/s	250 kb/s
Nominal range	up to 50 m	10 - 100 m
Nominal TX power	0 dBm	–25 to 0 dBm
Number of RF channels	79	25 (16 in 2.4 GHz)
Channel bandwidth	2 MHz	0.3/0.6; 2 MHz
Modulation	GFSK	O-QPSK
Spreading	FHSS	DSSS
Basic cell	Piconet	Star
Extension of the basic cell	None	Cluster tree, mesh
Max number of cell nodes	>65000	>65000
Data protection	16-bit CRC	16-bit CRC
Connectivity	Supported by Bluetooth V4.0+ devices	Dedicated devices
Interference avoidance	Adaptive frequency hopping scheme	Dynamic channel selection
Current consumption (TX, 0 dBm output power)	TI CC2541: 18.2 mA	TI CC2530: 29 mA
Current consumption (RX)	TI CC2541: 17.9 mA	TI CC2530: 24 mA
MAC design	Mostly TDMA	Flexible
Lowest current unit cost	TI CC2541F128RHAR: $2.38	TI CC2530F128RHAR: $3.97

Table 1. The comparison chart of Bluetooth Low Energy and ZigBee.

Even though the packet overhead is not considered yet, the normalized energy consumption of BLE would be smaller than ZigBee. Furthermore, because of the difference in their data rate, the transmission delays with BLE are smaller, and the delay performance can be very important for certain delay-sensitive vehicular applications.

Regarding the energy consumption, the current consumption of a BLE compliant chip is comparable to the current consumption of a ZigBee compliant chip. For example, the current consumption of a Texas Instruments CC2541 BLE compliant chip is 17.9 mA and 18.2 mA for receiving (RX) and transmitting (TX), respectively [13]. On the other hand, the current consumption of a Texas Instruments CC2530 ZigBee compliant chip is 24 mA and 29 mA for RX and TX, respectively. However, the maximum data rate of ZigBee is 250 kb/s, while BLE's is 1 Mb/s. Even though the packet overhead is not considered yet, the normalized energy consumption of BLE would be smaller than ZigBee. Furthermore, because of the difference in their data rate, the transmission delays with BLE are smaller, and the delay performance can be very important for certain delay-sensitive vehicular applications (such as safety applications).

Regarding reliability and robustness, BLE employs an adaptive frequency hopping scheme to combat coexistence and fading problems, while ZigBee employs dynamic frequency selection. Under interference, BLE can dynamically update the frequency hopping sequence to exclude the channels with interference during active communications. ZigBee, on the other hand, selects a clearer channel before the communication starts, and then it sticks to the selected channel. Although ZigBee can choose to change channels periodically, it is still not as dynamic as BLE. As a result, BLE could be more sustainable over transient interference. We reported that when no interference exists in a car, the performance of BLE and ZigBee for IVWSNs is comparable. However, if strong WiFi interference is introduced, BLE can provide better performance than ZigBee [8].

Compared to BLE, ZigBee provides greater flexibility in terms of network topology and MAC design. For instance, the basic topology of a ZigBee network is star, but it also supports cluster trees or mesh. On the other hand, BLE only supports piconets (and scatternets if Bluetooth version 4.1 is used) in connection mode, which follows a star topology. This, however, is not a problem for communications between ECUs and sensors, which typically have a star topology. One can consider the ECU as the master device in a piconet, and the sensors as the slave devices. The ECU (i.e. the master device) coordinates the communications of sensors in the piconet. Regarding the MAC design, since ZigBee applies direct sequence spread spectrum, although the standard MAC protocols of ZigBee employ CSMA and TDMA, ZigBee can also use a large number of customized MAC protocols based on CSMA, TDMA, FDMA, CDMA, or a combination thereof. However, BLE can only use time-division (or reservation-based) MAC protocols due to the nature of frequency hopping spread spectrum. Therefore, for an IVWSN based on BLE, it is necessary to carefully design a scheduler for each piconet in order to accommodate the requirements of each sensor/application in the piconet.

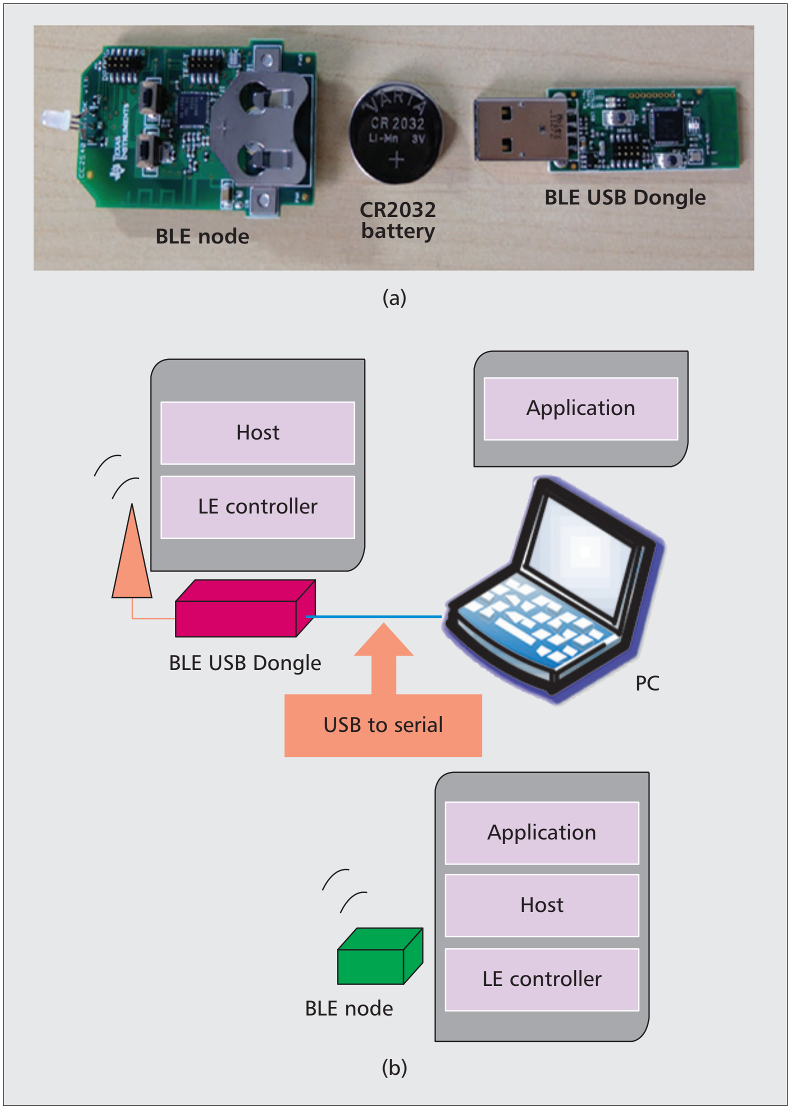


Figure 3. The Bluetooth Low Energy experimental platform: a) Texas Instruments CC2540 Mini Development Kit; b) system diagram.

Example Implementation

Experimental Platform

To show the utility of the BLE technology in vehicular applications, an example implementation was also carried out and experiments were performed with this implementation. The experimental platform used in this article is based on the Texas Instruments CC2540 Mini Development Kit [14]. The Texas Instruments CC2540 is a single-chip BLE solution that is capable of executing the BLE protocol stack and applications with a built-in 8051 microcontroller. The development kit includes a BLE node and a USB dongle, as shown in Fig. 3a. The BLE node is powered by a CR2032 coin battery. The architecture of our experimental platform is depicted in Fig. 3b. The USB dongle is connected to a PC with a USB to serial link. On the USB dongle, there are the host, an LE controller, and an adaptation layer that serves as the interface between the host and the PC. The application layer and a serial port interface are implemented on the PC. On the BLE node, there are the application layer, the host, and an LE controller. Note that in the real automotive platform, the application layer will be implemented on an ECU (instead of a PC), and it can use a universal asynchronous receiver/transmitter (UART) link to communicate with the CC2540 BLE chip.

A Passive Keyless Entry System

A passive keyless entry system refers to a vehicle that can detect the key in its proximity and unlock itself (or unlock when the user pulls the door handle) when the key appears within a certain range from the vehicle. Several car manufacturers currently provide similar features on their production vehicles. However, in many of the current implementations (which usually use a low-frequency RF to detect the transducer on the key fob), the current consumption of the system on the vehicle could be high, for example, around 700 mA in some GM cars. To prevent draining of the battery, the system has to enter the sleep mode when it is idle, and it incurs undesirable long latency when the system is reactivated. To address the high current consumption and the long latency issues and provide a solution with lower cost, as a proof-of-concept, we have designed a passive keyless entry system based on the proposed BLE IVWSN platform.

The test vehicle used in the experiment is a 2009 Cadillac STS. Two BLE nodes represent the BLE-enabled keys, and the USB dongle along with a PC is installed on the test vehicle to represent a lock control system on the vehicle. The keys are programmed as BLE peripheral devices. After powering on, the keys periodically send out advertisements with authentication information, and the keys will accept the connection if the connection request from the central device carries the correct pass code. On the vehicle, the USB dongle is programmed as a BLE central device, and its behavior is controlled by the application implemented on the PC. On the PC, there are three components in the application layer (Fig. 4a). One of the components is the connection manager, which initiates and maintains the BLE connections to the keys; the other component is the RSSI handler, which monitors the RSSI measurements of the packets from the keys and determines if the car should be unlocked. The third component is a serial interface for communicating with the USB dongle.

The flow chart of the connection manager is shown in Fig. 4b. The connection manager maintains a valid key list and an active key list. The valid key list is pre-defined and should be preprogrammed by the car manufacturer. According to the valid key list, the connection manager scans for advertisements from those valid keys and handles the BLE connections to them. If any valid key is discovered, the connection manager will initiate the connection to the key and add the key to the active key list. After the BLE connection is established, the RSSI measurements are taken during each connection event. In each connection event, the central device sends a packet to the key, and then the key sends another packet back to the central device. The RSSI measurement of the latter packet is collected by the RSSI handler. According to the active key list, the RSSI handler collects the RSSI measurements from all of the active keys and determines if the system should unlock the doors when the user pulls the door handle.

As shown in Fig. 5 the position of the key can be categorized into three regions based on the RSSI measurements. In region (**a**), the key is out of the BLE communication range (e.g. around 25 m when the transmission power is 0 dBm); in region (**b**), the key has an active connection to the central device; in region (**c**), the key has an active connection, and the RSSI of the packets from the key is larger than a predefined threshold (e.g. –55 dBm). Only when a valid key is in region (**c**) will the system unlock the doors of the vehicle when the user pulls the door handle. Also, when there is no active key in the range for more than a certain period of time (e.g. 30 seconds), the system can choose to lock the car.

The design was evaluated under the test case that the driver with the key walks toward the vehicle and pulls the door handle, and then walks away from the vehicle 50 times. The system could correctly unlock the car every time the driver pulled the handle and the response time experienced by the driver is negligible. This system can be easily integrated to the BLE IVWSN in future vehicles, thus providing a low-cost, low-latency, and highly efficient solution for a passive keyless entry system.

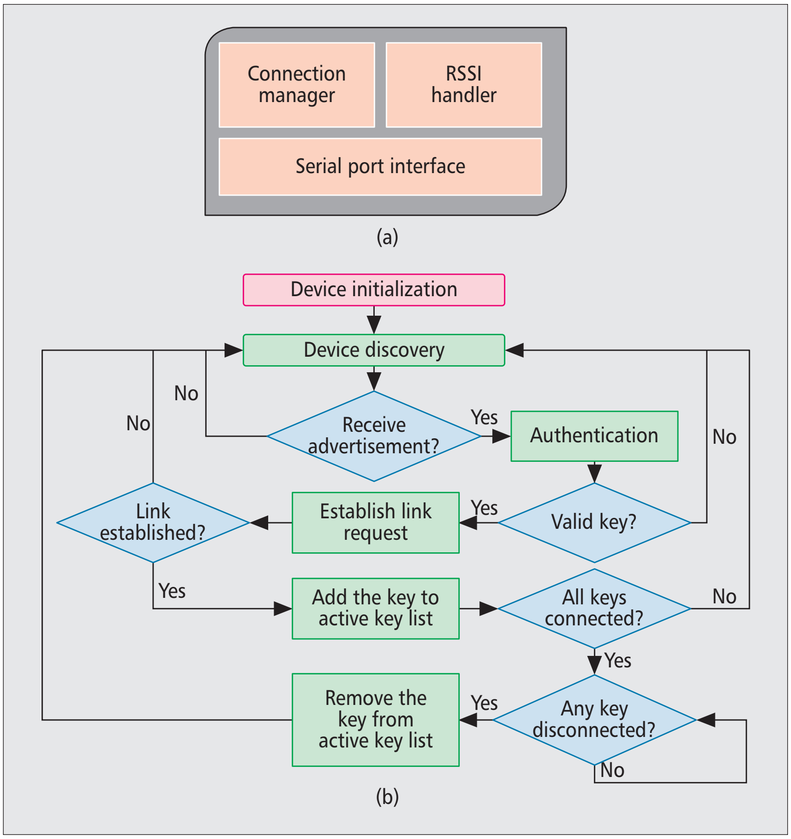


Figure 4. The application components of the passive keyless entry system: a) the components of the application layer on the central device; b) the state diagram of the connection manager component.

Discussion

In this article we have demonstrated an IVWSN experimental platform based on BLE technology. There are two different options for implementing the BLE IVWSNs in production vehicles. In addition to replacing wired sensors with BLE sensors, the first way to set up the network is to install several stand-alone BLE central devices and attach them to the vehicle bus. The BLE central devices would serve as gateways between BLE sensors and the ordinary ECUs. The main advantage of this approach is the fact that changes to the existing architecture and other components would be minimal. However, the total cost of the related components in one vehicle would be higher. The other option is to add a BLE compliant chip or daughter board into multiple ECUs, so that the ECUs can directly communicate with BLE peripherals. However, this approach involves changes to the current ECUs and the initial cost and the effort needed to make such changes will be larger. Also, the positions of the ECUs and the placement of BLE antennas will be additional important design issues.

The other major issues are the MAC design and the channel capacity of the system when there are multiple BLE master devices existing in a single vehicle. As mentioned in the previous sections, BLE supports mainly time-division MAC protocols due to the nature of its PHY layer characteristics. Therefore, for the deployment of IVWSNs in a production vehicle, it is critical to calculate a schedule for all the sensors and ECUs to follow in order to achieve maximum performance and minimize interference. We are currently investigating the underlying mechanism for designing such a scheduler.

It is important to note that BLE technology is fully equipped to protect the privacy and security of the communications. Encryption in BLE uses Advanced Encryption Standard in the counter with Cipher Block Chaining — Message Authentication Code Mode (AES-CCM) cryptography, and multiple keys are generated by the host for data and device authentication. BLE also supports a privacy feature that can change the Bluetooth device address on a frequent basis to prevent an

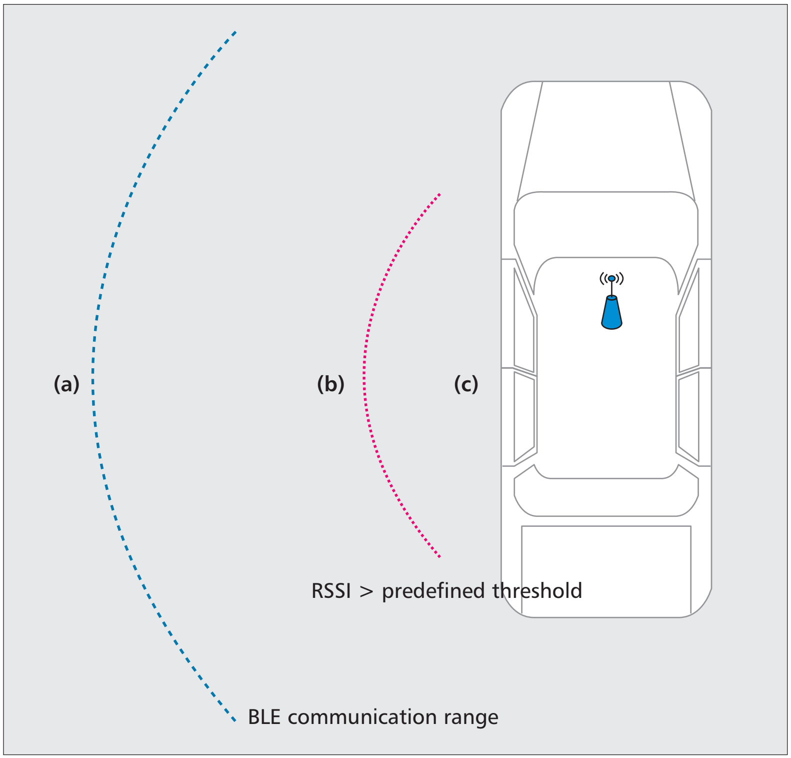


Figure 5. The operation of the Passive Keyless Entry System based on BLE.

LE device from being tracked by eavesdroppers. However, since one of the design goals of BLE is to keep the cost and complexity of a slave device to a minimum level, the association modes of BLE are not as sophisticated as those in BR/EDR. It has been reported that the key exchange during association could be compromised under certain circumstances [15]. For IVWSNs, since all of the devices are pre-installed in the vehicle, one potential solution is to pre-define and store the cryptographic keys in the ECUs and sensors. Future work should look into different ways of further enhancing the security of the system.

An important concern about the wireless sensors is the battery life. According to [16], the average current consumption of a BLE connection event is 10.655 mA, and the average duration is 2.348 ms. The current consumption during the sleep state is 0.9 μA. If the connection interval of the system is 2 seconds, then the average current consumption can be calculated as:

$$\begin{aligned} I_c &= (10.655 \text{ mA} \times 2.348 \text{ ms} + 0.9 \text{ mA} \\ &\quad \times \; 1997.652 \text{ ms})/2000 \text{ ms} \\ &\cong 0.013 \text{ mA} \end{aligned}$$

The typical capacity of a CR2032 coin battery is 230 mAh, so the estimated battery life is:

$$\begin{aligned} T_b &= 230 \text{ mAh}/I_c \\ &\cong 230 \text{ mAh}/0.013 \text{ mA} \\ &\cong 17692 \text{ hours} \cong 737 \text{ days} \cong 2 \text{ years} \end{aligned}$$

Therefore, if the connection interval is 2 seconds and the BLE sensor node is connected all the time, its estimated battery life can be up to 2 years, which is quite respectable for most vehicular applications.

Conclusion

In this article we have shown that Bluetooth Low Energy technology is an excellent choice for intra-vehicular wireless sensor networks (IVWSNs). An in-depth comparison between Bluetooth Low Energy and ZigBee in the context of IVWSNs is also provided and pros and cons of the two options are highlighted. Furthermore, we have reported an example application (namely, a passive keyless entry system) to demonstrate a use case for the Bluetooth Low Energy experimental platform for intra-vehicular wireless communications. The main motivation for implementing a passive keyless entry system based on Bluetooth Low Energy is to reduce the response time, power consumption, and the cost of the existing system. Overall, our results show that Bluetooth Low Energy is a promising and viable wireless technology for IVWSNs and certain automotive applications that require low-power and low-cost solutions.

References

[1] M. Ahmed *et al.*, "Intra-Vehicular Wireless Networks," *Proc. IEEE Globecom Workshops*, Nov. 2007.
[2] O. K. Tonguz and G. Ferrari, *Ad Hoc Wireless Networks: A Communication-Theoretic Perspective*, John Wiley & Sons, 2006.
[3] H.-M. Tsai *et al.*, "Zigbee-Based Intra-Car Wireless Sensor Networks: A Case Study," *IEEE Wireless Commun.*, vol. 14, no. 6, Dec. 2007, pp. 67–77.
[4] C. U. Bas and S. C. Ergen, "Ultra-Wideband Channel Model for Intra-Vehicular Wireless Sensor Networks Beneath the Chassis: From Statistical Model to Simulations," *IEEE Trans. Veh. Technol.*, vol. 62, 2013, pp. 14–25.
[5] "Bluetooth Core Version 4.0 specification," June 2010; available: https://www.bluetooth.org/Technical/Specifications/adopted.htm.
[6] "Bluetooth Core Version 4.1 specification," Dec. 2013; available: https://www.bluetooth.org/Technical/Specifications/adopted.htm.
[7] A. R. Moghimi *et al.*, "Characterizing Intra-Car Wireless Channels," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, Nov. 2009, pp. 5299–5305.
[8] J.-R. Lin, T. Talty, and O. K. Tonguz, "An Empirical Performance Study of Intra-Vehicular Wireless Sensor Networks Under Wifi and Bluetooth Interference," *Proc. IEEE Global Communications Conf.*, Dec. 2013.
[9] H.-M. Tsai *et al.*, "Zigbee-Based Intra-Car Wireless Sensor Network," *Proc. IEEE Int'l. Conf. Commun.*, 2007, pp. 3965–71.
[10] T. ElBatt *et al.*, "Potential for Intra-Vehicle Wireless Automotive Sensor Networks," *Proc. Sarnoff Symposium*, March 2006, pp. 1–4.
[11] ZigBee Alliance, "ZigBee 2007 specification," 2007; available: http://www.zigbee.org/Specifications.aspx.
[12] J.-S. Lee, Y.-W. Su, and C.-C. Shen, "A Comparative Study of Wireless Protocols: Bluetooth, UWB, Zigbee, and Wi-Fi," *Proc. 33rd Annual Conf. IEEE Industrial Electronics Society* (IECON), Nov. 2007, pp. 46–51.
[13] Texas Instruments, "CC2541 RF Transceiver Datasheet," June 2013; available: http://www.ti.com/lit/ds/symlink/cc2541.pdf
[14] ——, "CC2540 Mini Development Kit;" available: http://www.ti.com/tool/cc2540dk-mini.
[15] M. Ryan, "Bluetooth: With Low Energy Comes Low Security," *Proc. 7th USENIX Conf. Offensive Technologies*, USENIX Association, 2013, pp. 4–4.
[16] Texas Instruments, "Measuring Bluetooth Low Energy Power Consumption," *Application Note AN092*, August 2010.

Biographies

Jiun-Ren Lin is a Ph.D. candidate in the Electrical and Computer Engineering Department of Carnegie Mellon University (CMU). He received his B.S.E. in computer science and information engineering from National Chiao Tung University in 2004 and his M.S. in computer science and information engineering from National Taiwan University in 2006. He is a member of the General Motors–Carnegie Mellon University Collaborative Research Laboratory working on the design of intra-vehicular wireless sensor networks. His research interests include computer networks, wireless networks and communications, personal communications systems, vehicular networks, and performance evaluation.

TIMOTHY TALTY received his B.S.E.E. from Tri-State University, Angola, Indiana, in 1987, and his M.S. and Ph.D. from the University of Toledo, Ohio, in 1990 and 1996, respectively. From 1982 to 1988 he was employed as a civilian with Naval Sea Systems Command, U.S. Navy, Arlington, Virginia. He joined Ford Motor Company in 1993, and worked on wireless channel modeling and concealed antenna systems development. He joined the EECS Department of the United States Military Academy, West Point, New York, as an assistant professor in 1997, where he conducted research on embedded antenna systems and high-speed Sigma-Delta converters. In 2001 he joined General Motors Corporation, Warren, Michigan, where he is currently a technical fellow working in the areas of wireless sensors and networks.

OZAN K. TONGUZ is a tenured full professor in the Electrical and Computer Engineering Department of Carnegie Mellon University (CMU), Pittsburgh, Pennsylvania. He currently leads substantial research efforts at CMU in the broad areas of telecommunications and networking. He has published approximately 300 papers in IEEE journals and conference proceedings in the areas of wireless networking, optical communications, and computer networks. He is the author (with G. Ferrari) of the book *Ad Hoc Wireless Networks: A Communication-Theoretic Perspective* (Wiley, 2006). He is the founder, president, and CEO of Virtual Traffic Lights, LLC, a CMU spinoff that was launched in December 2010, which specializes in providing solutions to several transportation problems, such as safety and traffic information systems, using vehicle-to-vehicle and vehicle-to-infrastructure communications paradigms. His current research interests include vehicular ad hoc networks, wireless ad hoc and sensor networks, self-organizing networks, smart grid, bioinformatics, and security. He currently serves or has served as a consultant or expert for several companies, major law firms, and government agencies in the United States, Europe, and Asia.

NETWORK AND SERVICE MANAGEMENT

George Pavlou

Juergen Schoenwaelder

This is the 18th issue of the series on Network and Service Management, which is published twice a year. Until 2012 it was published in July and December, but since 2013 it is published in January and July. The series provides articles on the latest developments in this well established discipline, highlighting recent research achievements and providing insight into both theoretical and practical issues related to the evolution of the discipline from different perspectives. The series provides a forum for the publication of both academic and industrial research, addressing the state of the art, theory, and practice in network and service management.

One important change to the series during the last year is that Prof. Aiko Pras, who was a co-editor of the series from the very first issue in October 2005 until January 2014, has decided to step down. We are delighted to welcome on board Prof. Juergen Schoenwaelder of Jacobs University of Bremen who is very well known in the network and service management research community, having produced a number of seminal contributions over the years (http://cnds.eecs.jacobs-university.de/users/schoenw/). Prof. Schoenwaelder has actively contributed to previous issues as well as this one, and is expected to put his own stamp on the series in the years to come.

An important event for both network/service management but also for the networking research community as a whole is the appointment of Prof. Aiko Pras of the University of Twente, Netherlands, as Chair of the whole IFIP TC6 Communication Systems, taking over from Prof. Guy Leduc of the University of Liege, Belgium. Prof. Pras has also been Chair of IFIP 6.6 Management of Networks and Distributed Systems since 2008. As already mentioned, Prof. Pras was also a co-editor of this series since its inception, having stepped down exactly a year ago.

The key annual event in this area, the IEEE/IFIP Network Operations and Management Symposium (NOMS 2014), was held May 5–9 in Krakow, Poland (http://ieee-noms.org/2014/). The second key annual event in this area was the 10th instance of the IFIP/IEEE International Conference on Network & Service Management (CNSM 2014), which has become another flagship event complementing IM and NOMS. This took place November 17–21 in Rio de Janeiro, Brazil; http://cnsm-conf.org/2014/. This year is the turn of the IFIP/IEEE Integrated Management Symposium (IM 2015), which will take place May 11–15 in Ottawa, Canada; http://ieee-im.org/2015/. Note also the 1st IEEE Conference on Network Softwarization (NetSoft 2015) to be held in April 13–17 in London, United Kingdom; http://sites.ieee.org/netsoft/. Finally, the European Conference on Autonomous Infrastructure Security & Management (AIMS), which is supported by the FLAMINGO Network of Excellence (see below), will be held June 22–29 in Ghent, Belgium; http://www.aims-conference.org/2015/.

Finally, it should be mentioned that the European Network of Excellence FLAMINGO on the Management of the Future Internet has already been running for two years and is expected to run for another two. One of the things currently worked on in the project is a revised network and service management taxonomy. This was discussed during:

- The CNOM meeting at IEEE GLOBECOM 2014
- The CNOM and IFIP 6.6 meeting at NOMS 2014
- The NMRG meeting in Vancouver 2014
- The EC stakeholders consultation (session on network management) in September 2014, Brussels

For an overview of the project's activities see http://fp7-flamingo.eu/.

We have again experienced an overwhelming interest in the 18th issue, receiving 18 submissions in total. Each of the articles received at least three independent reviews. We finally selected three articles, resulting in an acceptance rate of 16.7 percent. It should be mentioned that the acceptance rate for all the previous issues has ranged between 14 and 25 percent, making this series a highly competitive place to publish. We intend to maintain our rigorous review process in future issues, thus maintaining the high quality of the published articles.

The first article, "Software Defined Networking: Management Requirements and Challenges" by Wickboldt, de Jesus, Isolani, Both, Rochol, and Granville, classifies and discusses how to manage networks based on the emerging SDN paradigm, presenting some of the requirements, describing current proposals, and highlighting relevant challenges.

The second article, "A Recommender System Architecture for Predictive Telecom Network Management" by Zaman, Hogan, van der Meer, Keeney, and Muntean, presents a predictive model for telecom network management that avoids reactive operation by observing event sequences and recommending preemptive actions to prevent problems before they occur.

Finally, the third article, "Management Architecture for Location-Aware Self-Organizing LTE/LTE-A Small Cell Networks" by Fortes, Aguilar-Garcia, Barco, Barba, Fernandez-Luque, and Fernandez-Duran, proposes a novel architecture for next-generation cellular networks at indoor scenarios, supporting self-organizing functions based on the knowledge of the user equipment location.

We hope that readers of this issue again find the articles informative, and we will endeavor to continue with similar issues in the future. We would finally like to thank all the authors who submitted articles to this series and the reviewers for their valuable feedback and comments on the articles.

BIOGRAPHIES

GEORGE PAVLOU (g.pavlou@ucl.ac.uk) is a professor of communication networks in the Department of Electronic and Electrical Engineering, University College London, United Kingdom, where he coordinates research activities in networking and network management. He received a Diploma in engineering from the National Technical University of Athens, Greece, and M.Sc. and Ph.D. degrees in Computer Science from University College London. His research interests focus on networking and network management, including aspects such as traffic engineering, quality of service management, policy-based systems, autonomic networking, information-centric networking, and software-defined networks. He has been instrumental in a number of European and U.K. research projects that have produced significant results with real-world uptake, and has contributed to standardization activities in ISO, ITU-T, and the Internet Engineering Task Force (IETF). He has been the Technical Program Chair of several conferences, and in 2011 he received the Daniel Stokesbury award for "distinguished technical contribution to the growth of the network management field."

JÜRGEN SCHÖNWÄLDER (j.schoenwaelder@jacobs-universiy.de) is a professor of computer science at Jacobs University Bremen, Germany, where he leads the Computer Networks and Distributed Systems research group. He received a doctoral degree from the Technische Universität Braunschweig, Germany. His research interests include network management, distributed systems, network measurements, embedded networked systems, and network security. He is an active member of the IETF, where he has edited more than 30 network management related specifications and standards. He co-chaired the ISMS working group of the IETF and currently serves as co-chair of the NETMOD working group. Previously, he chaired the Network Management Research Group of the Internet Research Task Force. He has been involved in several European research projects and served in various roles for IEEE and IFIP sponsored conferences. He currently serves on the Editorial Boards of the Springer *Journal of Network and Systems Management* and the Wiley *International Journal of Network Management*. Previously, he served on the Editorial Board of *IEEE Transactions on Network and Service Management*.

Software-Defined Networking: Management Requirements and Challenges

Juliano Araujo Wickboldt, Wanderson Paim de Jesus, Pedro Heleno Isolani, Cristiano Bonato Both, Juergen Rochol, and Lisandro Zambenedetti Granville

The authors are with Federal University of Rio Grande do Sul (UFRGS).

ABSTRACT

SDN is an emerging paradigm currently evidenced as a new driving force in the general area of computer networks. Many investigations have been carried out in the last few years about the benefits and drawbacks in adopting SDN. However, there are few discussions on how to manage networks based on this new paradigm. This article contributes to this discussion by identifying some of the main management requirements of SDN. Moreover, we describe current proposals and highlight major challenges that need to be addressed to allow wide adoption of the paradigm and related technology.

INTRODUCTION

Software-Defined Networking (SDN) is a network paradigm usually characterized by three fundamental aspects:

- A clear separation of network forwarding and control planes.
- The abstraction of the network logic from hardware implementation into software.
- The presence of a network controller that coordinates the forwarding decisions of network devices.

Given that software, in SDN, can be more easily coded, deployed, and executed, SDN turns out to be a very disruptive technology that better promotes network innovation. In addition, SDN has been grabbing the attention of both industry and academia, and has experienced strong support by major Internet players (e.g. Google, Cisco, NEC, Juniper) and standardization bodies (e.g. ONF and IETF). Today SDN is a driving force in the field of computer networks.

Much has been discussed about SDN and network management, mostly from the perspective of where SDN is taken as a management tool [1]. There are several benefits, from the traditional network management point-of-view, in adopting SDN because it simplifies or even solves critical management tasks. For example, because SDN devices need to be registered or discovered by the network controller in order to establish a communication path between control and forwarding planes, network discovery, a traditional network management task, is intrinsically solved.

In this article we take a different perspective on SDN and network management. Although SDN does solve some classical management problems, it also creates new ones. This fact turns out to be a traditional "meta-problem" in the area of network management: every time a new network technology is introduced, the importance of its management is generally underestimated. When the technology matures and is widely deployed, network management arises as a real need. However, because management did not evolve together with the newly introduced networking technology, frequently network management becomes just a technology patch. As such, addressing SDN management is imperative to avoid patching SDN later.

In this article we address SDN management requirements to foster the adoption and development of the topic. We investigate to what extent current proposals for SDN management suffice, review how research in the area has evolved, and finally highlight management challenges that arise from the current picture of SDN management. The remainder of this article is organized as follows. In the next section we present the concepts and evolution of SDN. We then define management requirements of SDN. After that, current proposals for SDN management are presented. Next, challenges on SDN management are discussed. Finally, we summarize the article, presenting our final remarks.

SDN: EVOLUTION AND CONCEPTS

This section presents an evolution timeline and a review of SDN concepts. It is important to emphasize that SDN concepts are still under formation. Nevertheless, from the interest and especially given the significant investments made by the industry, one can safely state that this new network paradigm has a lot of potential to succeed. From one perspective, it is

0163-6804/15/$25.00 © 2015 IEEE

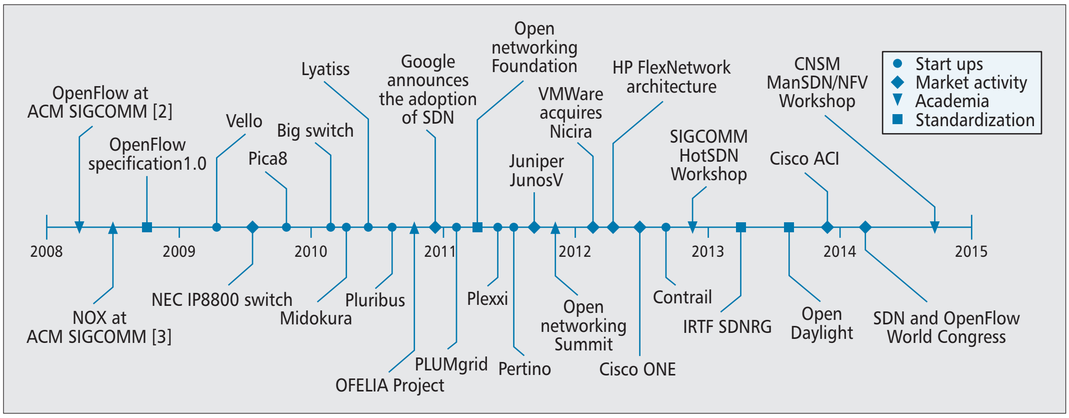


Figure 1. Major events on the SDN timeline over the last seven years.

Academia has played an important role in this evolution, early during SDN conception, during recent last years maturing the concepts, and regularly stimulating discussions through research projects, such as the development of the OFELIA testbed in Europe.

impressive how SDN is able to foster innovation in networks at a level that has not been seen in the last decade. On the other hand, companies are usually too concerned with time to market rather than providing conceptually elegant solutions. Particularly, the main issue we envision in this scenario is that the majority of efforts are focused on providing solutions and services over SDN networks, whereas network management is not given much attention. However, as widely recognized, management must not be an afterthought [2].

Evolution Timeline

In order to observe the evolution of SDN from the perspective of both industry and academia, we present a timeline containing some of the major events in this area (Fig. 1). Our investigation begins in the earliest phase of OpenFlow [3], arguably the most important SDN implementation today. We did not include pre-SDN proposals, such as ETHANE, because the concept was not then referred to as SDN yet. Moreover, that period has already been discussed [4]. OpenFlow was first introduced in a white paper in March 2008, from the network research group of Stanford University and collaborators. Given the importance of the work, the paper appeared as an editorial note in *ACM SIGCOMM Computer Communication Review* one month after that. The objective back then was simple: to enable researchers to innovate, and try out their proposals, within a campus network. The first time the concept of "network operating system" appeared in this context was through the proposal of the NOX controller [5]. NOX represents the first attempt to enable the development of software to control OpenFlow-based networks. In the context of SDN, such a development task may turn out to be a common responsibility among software developers and network administrators.

Right after launching the first specification of OpenFlow, with official vendor support,[1] at the end of 2009, there was much excitement in the networking market. NEC was the first to announce a commercial switch with native OpenFlow support, in April 2009. Also, start-up companies, such as Pica8, Big Switch, Plexxi, and Vello began to pop out offering SDN-ready solutions. In 2011 some Internet giants, such as Google, announced the adoption of SDN inside their data centers and backbones. In the same year a few standardization efforts also started. ONF is basically dedicated to promote and evolve the concept of SDN by standardizing the OpenFlow technology. Near 2012 other major players such as Cisco, Juniper, Hewlett-Packard, and VMWare rushed to put themselves in a leading position in that promising but not yet matured networking paradigm. In the beginning of 2013 the IRTF started the Software-Defined Networking Research Group (SDNRG), while the Linux Foundation launched the Open Daylight project. All these standardization efforts are supposed to be vendor independent, although most of the aforementioned companies actually embody the majority of the working groups and committees. Academia has played an important role in this evolution, early during SDN conception, in recent years maturing the concepts, and regularly stimulating discussions through research projects, such as the development of the OFELIA testbed in Europe, and conferences, such as the Open Networking Summit, the SIGCOMM HotSDN Workshop, and CNSM ManSDN/NFV Workshop.

Revisiting SDN Concepts

SDN and its properties have been defined by several authors in the last few years. SDN is commonly defined as a new networking paradigm in which the forwarding plane is decoupled from the control plane [5]. In addition, most of the "intelligence" of an SDN network is implemented inside a controlling entity, while network devices (switches) are only simple packet forwarding boxes. In fact, this definition is biased by the association of the concepts of SDN with its main enabling technology, that is, OpenFlow. Separating the forwarding and control planes and adding a logically centralized controlling entity are two characteristics that have often appeared in other types of networks.[2,3] This rather vague definition leaves the doors open for divergent SDN instantiations. As a consequence, many emergent network solutions propose a variety of concepts that have significant overlap with this definition, such as Software-Friendly Networks, Software-Driven Networks, Deeply Programmable Networks, and Network Functions Virtualization.

[1] First specification available online is 0.8.9 from December 2008, although the first for official vendor support was 1.0.0, released in the last day of 2009.

[2] IETF RFC 3654 — http://tools.ietf.org/html/rfc3654 3

[3] GPP TS 23.002 GSM Network architecture — http://www.3gpp.org/DynaReport/23002.htm.

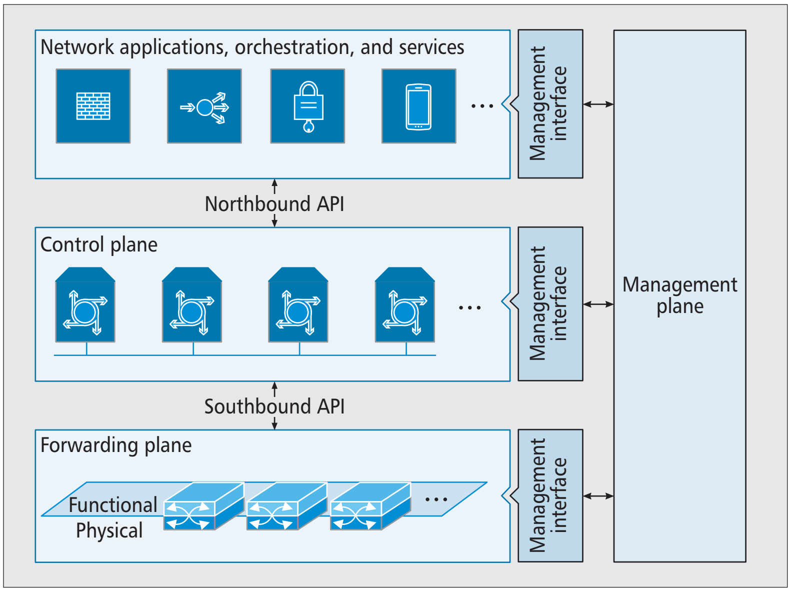


Figure 2. High-level conceptual architecture of SDN.

We look at SDN from a slightly different angle than many other authors [1, 3, 5]. We consider that SDN is essentially about abstracting the logic of traditional networks from within the fixed hardware implementation, thus raising it to a higher level defined by software. Separating the logic of the forwarding and control planes and laying down a network controller entity are also important concepts concerning architectural organization. However, most fundamentally, software within all planes needs to dictate how traffic is handled in the network.

Observing the state-of-the-art in SDN, we are able to depict a general architecture composed of four planes and three interfaces, as depicted in Fig. 2. Observing the architecture from a bottom-up approach, the forwarding plane is split into "functional" and "physical." The functional part will always be a collection of software functions to be executed in forwarding devices. This functional part is the reason why we call this the "forwarding plane" instead of "data plane," as many other authors do. From our point-of-view, the functional part does not only hold data, but it must allow the use of software to perform forwarding tasks, employing tables, trees, queues, or any other internal data structure. The remaining elements in this plane are strictly physical, such as I/O ports, memory, processor, and storage.

The control plane is originally meant to compile higher-level decisions and enforce the necessary configurations into forwarding devices. Elements in the control plane should at least execute requests coming from the plane above. Commonly, these elements also include internal logic to handle network events and recover from failures. Environment-specific demands might dictate how this plane should be designed, defining the best strategies for locating and distributing elements. On the top level of the SDN architecture, network applications, orchestration functions for business and network logic, as well as high-level network services are found, for example, load balancers and firewalls. We take a generic approach, including in the architecture a management plane. This plane rarely appeared in early SDN proposals, although it is being gradually considered in recent specifications.[4] We consider the management plane a fundamental part of the architecture to organize the implementation of operations, administration, and management (OAM) functions in SDN.

To allow information to flow between the management plane and other planes, management interfaces are in place. These interfaces allow settings to be enforced in network devices and information to be retrieved back into the management plane, for example, to be reported to the network administrator. In addition, two other important interfaces exist: a southbound API that allows the forwarding plane to communicate with the control plane, and a northbound API that abstracts control plane functions to network applications at the top level.

Management Requirements of SDN

Given the current noticeable paradigm shift to SDN and its new concepts, the development of a management plane as presented in Fig. 2 encompasses satisfying new management requirements. In this section we discuss some of the most relevant requirements considering all planes of the SDN architecture. We also summarize such requirements in Table 1, providing a comparison with traditional networks and management activities examples.

Bootstrap and Configuration

In traditional networks the logic of the forwarding and control planes is confined inside each network device according to a well defined set of standardized protocols. With the separation of planes, as SDN promotes, the need to bootstrap the communication between the forwarding and control planes becomes a basic requirement. Configuring this communication can be particularly complex, considering that both planes can operate under protocols defined by software. Moreover, software changes in any plane may affect such communication directly. Ideally, new software-defined protocols should include management interfaces to enable proper bootstrap and configuration in these networks.

Availability and Resilience

As in any other network, SDN is susceptible to faults in physical or logical elements, for example, link failures or network software bugs. Particularly, with the decoupling of planes comes the possibility of eventual disconnection between the forwarding and control planes. Many approaches to SDN even consider placing the whole control plane inside a single node, in which case unavailability and resilience issues become even more critical because of a single point of failure. Despite the fact that forwarding devices are usually able to react to this issue, it is still important to manage whether the connection between planes is active and in accordance with the policies of the network.

[4] SDN Architecture 1.0 (June, 2014) — https://www.opennetworking.org/sdn-resources/sdn-library/technical-papers.

Management requirement	Traditional networks	Software-defined networking
Bootstrap and configuration	Set well known protocol parameters, track configuration changes	Configure customized and ever-changing software, setup forwarding and control plane connectivity
Availability and resilience	Configure alternative routes in case of link failure	Configure forwarding devices behavior in case of failure in the connection with control plane
Network programmability	Not required	Control versioning, coordinated deployment, and verification of network software
Performance and scalability	Bandwidth assignment and reservation, Quality of Service configuration and enforcement	Monitor performance of network applications, adjust connection quality between forwarding and control planes
Isolation and security	Control network access, prevent intrusion, spoofing and Denial of Service	Grant isolation to network applications, prevent eavesdropping and usurpation of control traffic
Flexibility and decoupling	Adjust the management of higher level protocols	Adapt management functions along with management interfaces, coordinate management information within planes or among management systems
Network planning	Assess capacity and performance needs, choose a network topology	Plan the disposition of controlling elements in relation to forwarding elements
Monitoring and visualization	Track resource utilization, identify outages and trigger alarms	Trace functional parameters of novel applications, visualize jurisdiction of network controllers

Table 1. Traditional versus software-defined networking management activities.

Network Programmability

In the daily life of SDN administrators, every new network software release or update must be consistently persisted over the forwarding and control plane implementations across the network. To provide proper support for network programmability management, tools to allow network administrators to control versioning, coordinated deployment, rollback, and verification of network software are required. Moreover, software to dictate network behavior can also be developed in a variety of abstraction levels (high or application level, mid or control plane level, and low or forwarding plane level). Tools and methods need to be designed to decompose these high/mid-level policies or commands down to low-level device configurations.

Performance and Scalability

In SDN, network hardware needs to be more generic to allow software to be written in high levels of abstraction. Moreover, the separation of planes itself implies extra communication requirements between planes, which may add delay to network traffic being forwarded. Part of the responsibility of performance assurance falls on the shoulders of software developers that need to design optimized software for networks. Another part resides on network administrator's effort to understand bottlenecks, tweak the correct parameters to optimize software-defined protocols, and choose the most efficient control and management models to be employed (e.g. centralized, distributed, or hierarchical).

Isolation and Security

In SDN not only the network traffic is shared among many different users and applications, but also the network logic itself is controlled remotely by custom software. New resource isolation and security management techniques must be investigated to guarantee isolation in the three SDN planes as well as in their communications. Specifically, one must ensure that control traffic (flowing between controllers and forwarding devices) is isolated, and that code written by one developer does not affect the code written by others (e.g. by defining conflicting rules or resource race conditions). As SDN becomes more popular, vulnerabilities will appear and soon enough people will start writing the first malicious code for these networks.

Flexibility and Decoupling

Interfaces are required to allow exchange of management information from/to forwarding, control, and application planes. Since the behavior of the network can now be defined by software, SDN management needs to be flexible enough to quickly adapt to new protocols written for all planes. For example, one could implement a novel protocol for Information Centric Networking (ICN), which routes information relying on a caching system inside network devices. Managing cache performance (e.g. miss rates, occupation, object sizes) is a function that should be present in the management plane to allow administrators to adjust this protocol-specific parameter to optimize the network.

Network Planning

Traditionally, administrators need to plan for both deployment and expansion of networks, which encompasses tasks such as defining capacity and performance needs and deciding whether and where in the topology the network will be segmented (on layers 2 and 3). These decisions will result into acquiring a set of network hardware boxes (e.g. routers, switches, and firewalls) that operate under well known standard protocols. Planning is not a requirement for the SDN management plane per se, but it impacts and influences the other requirements. In SDN, administrators now need to acquire a set of forwarding devices and possibly another set of controlling devices. On top of that, network topology and capacity planning can be influenced by the type of software one installs on the network, also requiring software acquisitions. Administrators are still able to choose from different vendors of network software, controlling devices, and forwarding devices. Positioning these elements across the network topology can directly influence the performance, resilience, and survivability. In addition, the management plane could assist the SDN administrator in making planning decisions through a set of planning tools deployed in the SDN management frontend.

Positioning these elements across the network topology can influence directly the performance, resilience, and survivability. In addition, the management plane could assist the SDN administrator in making planning decisions through a set of planning tools deployed in the SDN management frontend.

Monitoring and Visualization

For the physical part, monitoring and visualization requirements remain similar to traditional networks. The logical part is far more complex assuming that forwarding and control protocols can be completely redesigned. For example, in current IP networks it is easy to draw a reachability map by analyzing information of routing tables, even before traffic starts flowing. This analysis is only possible because forwarding behavior of IP routers is predictable. When the internal implementation of a protocol is not known in advance, or defined by software, that is no longer possible.

Current Management Proposals for SDN

In this section we briefly review recent proposals that deal with some of the requirements presented in the previous section. We check whether these requirements are addressed fully, partially, or not addressed at all by current efforts. The proposals are grouped according to the SDN architectural part, being described starting by the forwarding plane, followed by the control plane, application plane, and finally the communication between planes. Since most proposals do not include explicitly a management plane in their conceptual architectures, we left this out of the organization of this section.

Given the need to deploy and set up novel forwarding devices into a traditional network, administrators must configure and bootstrap these devices in order to have them operational. In the most basic and common scenario, the administrator would interact with a command line interface of the device and have it configured by performing many intricate proprietary commands. In the case of SDN forwarding devices, there are still a few parameters to be configured, such as the communication policy with the control plane (e.g. drop every packet when the control plane is unavailable or follow the last valid set of rules). In response to these necessities, ONF proposed the OpenFlow Management and Configuration Protocol (OF-Config).[5] This protocol is based on NETCONF and allows operational staff to assign controller(s) to switches, set ports up/down, configure queues, assign certificates for the communication with the control plane, set up tunnels, handle versioning, and retrieve device capabilities. OF-Config is already a significant step toward tackling forwarding plane-related management requirements, such as *bootstrap and configuration* and *availability and resilience*. On the other hand, this protocol is targeted specifically to OpenFlow networks; therefore, it is tied to the limited view of SDN employed by this technology (e.g. fixed forwarding plane and logically centralized control plane).

The loose coupling between the forwarding and control planes in SDN allows greater flexibility in relation to the design and organization of elements in the control plane. Thus, the SDN control plane can be engineered with the strategy that best addresses the demands of users, such as availability, performance, and security demands. Proposals such as NOX [5], ONIX [6], HyperFlow [7], and Kandoo [8] are examples of different control plane implementations. While NOX adopts a simpler centralized strategy, ONIX, HyperFlow, and Kandoo employ distributed or hierarchical approaches where multiple controllers cooperate to control the underlying forwarding devices. These approaches are directly aligned with the *availability and resilience* requirement, and marginally with *performance and optimization*, and not aligned with *flexibility and decoupling*. To fit this last requirement, proposals should implement management interfaces as depicted in Fig. 2.

Network protocols and services, traditionally placed inside each network device, are now centralized in a loosely coupled way into the applications plane. Thus, tools previously employed to manage network services are incompatible with this new environment. Such tools must be redesigned under different requirements, such as *network programmability*, *isolation and security*, *monitoring and visualization*, and *flexibility and decoupling*. An attempt to address such requirements was led by RouteFlow [9], which allows traditional network protocols to run into virtualized network environments, although the routing decisions are still enforced in the underlying network using OpenFlow forwarding rules. Besides grouping network applications coherently, as proposed in RouteFlow, it is important to enable network administrators to manage applications.

SDN applications require high-level network abstractions, ideally presented through standardized interfaces, to communicate with elements in the control plane. A step toward meeting this requirement was given by SDN programming languages, such as Frenetic, Nettle, NetCore, Procera, and Pyretic [10]. These languages are predominantly declarative, but vary in objective, which includes efficiently expressing packet-forwarding policies, handling overlapping or conflicting rules, and improving horizontal scalability. From the operational point of view, SDN programming languages contribute to streamline the software development process, but to achieve *flexibility and decoupling*, developers must provide external interfaces to allow the execution of management tasks.

[5] OpenFlow Management and Configuration Protocol 1.2 (OF-Config 1.2) https://www.opennetworking.org/sdnresources/onf-specifications/openflow-config.

As discussed within the *availability and resilience* requirement, it is of fundamental importance to ensure the correct communication between the forwarding and control planes. FlowVisor [11] and its variations AdVisor [12] and VeRTIGO [13] are examples of tools that operate as proxies for delivering control information to different SDN controllers according to predefined rules. The employment of such tools allows network administrators to split the control logic of a network into several independent controlling devices. AdVisor, specifically, proposes the use of a management interface to control the rules used to forward control messages. Because of insufficiency of management interfaces and restricted management target, that is, only applied to communication between forwarding and control planes, these approaches are considered just partially aligned with the requirements of *flexibility and decoupling* and *monitoring and visualization*.

It is important to highlight that the aforementioned proposals have not been designed considering our list of management requirements, thus the mapping of the scope of each proposal to each requirement is not always direct. Still, one can now better observe how the state-of-the-art covers the listed management requirements.

Key Research Challenges in SDN Management

In this section we identify and discuss a non-exhaustive set of research challenges for SDN management. Instead of listing challenges that would cover the FCAPS (fault, configuration, accounting, performance, security) model, widely employed in network management in general, these challenges are a result of our analysis of the management requirements, and the coverage of proposals, discussed previously. Therefore, we highlight the main gaps found between requirements and proposals for SDN management.

From High-Level Rules to Network Configuration

Lower-level SDN planes introduce their own software abstractions and expose APIs to handle these abstractions to the upper-level planes. Every time the level of abstraction is raised, there is loss of information because not all low-level details are accessible through the defined APIs. When very high-level commands or rules come from the topmost planes of the SDN architecture, the lost information needs to be reconstructed or translated into a set of lower-level actions. To properly tackle the *bootstrap and configuration* requirement, processes for translating high-level rules into low-level configurations become a challenge. Although the vast literature on policy-based network management (PBNM) presents a solid base in this regard, further and SDN-focused research is required to tackle this challenge. We envision as a challenge, for the management plane of SDN, dealing with issues, such as: How is this translation of high-level rules (e.g. reduce latency, increase link redundancy, or save energy) realized in SDN-specific setups? How is the per-level rule refinement performed considering that the target infrastructure operates using SDN? Do SDN peculiarities generate conflicts that should be handled in high-level rules, and how should these conflicts be solved taking into account the low-level SDN details?

Autonomic and In-Network Management

Ideally, a management plane should concentrate all network management functions of an SDN. Nevertheless, since SDN allows software to dictate the behavior of network devices, it should be possible to place code inside these devices to allow them to react to network conditions and perform self-management functionalities. Autonomic and in-network management approaches offer the capability to migrate management functions, usually deployed in the management plane, to software running on devices of both forwarding or control planes. Autonomic and in-network are also important because they help addressing the requirement of *availability and resilience* in SDN. These approaches are generally considered in scenarios where the amount of devices or network connectivity conditions are unsuitable to maintain a frequent communication between devices and the management plane. In these scenarios, for instance, the implementation of fast failover capabilities (e.g. self-healing) directly in network devices becomes essential. Therefore, one important question to be answered is: What are the criteria to define where in the network or in which plane should autonomic loops be deployed and controlled?

Flexible Management Through Interfaces

In the SDN architecture, the management plane interacts with other planes through the use of interfaces. Adequate management interfaces are also required, for example, to enable a more integrated view of network resources with the system running on top of them, such as in cloud computing environments. To cope with the *flexibility and decoupling* requirement, the challenge of finding a systematic way of defining, using, and evolving management interfaces arises. Defining which management information flows where, when, and how encompasses answering a number of questions, such as: Who is the person responsible for writing the required interfaces, that is, the developer, the administrators, or both? Is it possible to make interfaces generic enough to be used or reused when a new software is developed and installed in networks? Can we take advantage of standards for interfaces already proposed or do we need new standards? Is there the need for an intermediary element to host these interfaces, such as a gateway?

Smart Network Planning

In SDN, administrators must consider three main aspects when planning a network infrastructure:

1. The physical separation of planes.
2. The software abstractions that drive network logic.
3. The existence of one or more controlling entities.

Lower level SDN planes introduce their own software abstractions and expose APIs to handle these abstractions to the upper level planes. Every time the level of abstraction is raised, there is loss of information because not all low-level details are accessible through the defined APIs.

In these times of "networking democracy", a crucial opportunity presents itself to addressing management requirements, and to avoid the recurrent mistake of patching management solutions after other concepts are already mature.

Regarding aspects 1 and 3, it is fundamental to consider that the number of elements in every plane can be arbitrary, as presented in Fig. 2. Moreover, considering aspect 2, choosing the right applications for networks becomes an essential and non-trivial task. In addition to being a requirement in SDN, planning is a challenge also because when virtualization is employed in SDN to support different networks on top of a single substrate, the amount of information the network administrator needs to handle is much larger in comparison with a non-virtualized network. Smart planning solutions become necessary to assist the network administrators with questions that were uncommon before the inception of SDN. Therefore, interesting questions need to be answered by SDN planning management solutions, such as: How many forwarding and controlling devices should be acquired to deploy the network? Where exactly should controlling devices be placed physically in relation to forwarding ones in the topology? How are control and management responsibilities organized (e.g. centralized, distributed, hierarchical)? How do software choices affect the physical topology? Are there compatibility issues among different software that need to coexist to form the deployed SDN?

Situational Management

In SDN the network conditions are much more dynamic because of facilitated, frequent software installation and changes. As a consequence, this is also more likely to create situational issues, which are not predicted in the design of management systems. To deal with unexpected and temporary conditions, such as debugging a newly installed software protocol, tools need to be available for fast creation of on-demand management applications taking advantage of information available from SDN planes and interfaces. Adequate situation management also encompasses fulfilling the requirement of *monitoring and visualization* of SDN [14]. Enabling situational management in SDN includes finding proper answers to: Which technologies are most suitable for creation of temporary management applications? Which techniques can be employed to combine information from multiple levels of the SDN architecture into on-demand management applications?

Conclusion

Despite being a relatively recent networking paradigm, the importance of SDN is already evidenced by the emergence of many start-up companies and foundations, the interest of researchers, and the support of major Internet players. Much has been recently discussed on taking SDN as a tool to simplify some classical network management issues. However, in this article we took a different perspective by analyzing the management necessities that did not exist before the inception of SDN.

Initially we revisited the discussions about the definition of SDN regarding concepts such as separation of planes and the actual implementation of a control plane. We approached SDN from a slightly different angle than many other authors, in which we emphasize the fact that SDN is essentially about abstracting network logic from hardware implementation to software. We also provided evidence that SDN currently overlaps with other emerging related concepts, such as Network Functions Virtualization and Software-Friendly Networks. Furthermore, we included in the SDN architecture a conceptual plane dedicated to implementing management functions, either in a centralized or distributed manner.

Aiming to encourage forthcoming proposals to take into account management concerns at the design phase, we discussed a non-exhaustive set of SDN management requirements. We kept our requirements irrespective to any technology and generic enough to remain valid over different SDN deployment designs. Moreover, we also discussed some of the most important investigations toward management of SDN already proposed and to which extent these investigations consider, directly or indirectly, one or more management requirements.

Finally, we established a set of challenges that we consider a fundamental contribution to encourage future investigations regarding SDN management. We envision mainly the resurgence of traditional network management concepts, such as autonomic/self-management and policy-based network management. Moreover, we believe in the empowerment of situational management, for example, based on mashup-oriented technologies. Most importantly, we understand that SDN represents a landmark: for the first time in decades we are witnessing computer network development happening outside private industry boundaries. In these times of "networking democracy" a crucial opportunity presents itself to address management requirements, and to avoid the recurrent mistake of patching management solutions after other concepts are already mature.

References

[1] H. Kim and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 114–19.

[2] J. Schonwalder *et al.*, "Future Internet = Content + Services + Management," *IEEE Commun. Mag.*, vol. 47, no. 7, July 2009, pp. 27–33.

[3] N. McKeown *et al.*, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, April 2008, pp. 69–74.

[4] N. Feamster, J. Rexford, and E. Zegura, "The Road to SDN: An Intellectual History of Programmable Networks," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 2, Apr. 2014, pp. 87–98.

[5] N. Gude *et al.*, "NOX: Towards an Operating System for Networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 3, July 2008, pp. 105–10.

[6] T. Koponen *et al.*, "ONIX: A Distributed Control Platform for Large-Scale Production Networks," *Proc. 9th USENIX Conference on Operating Systems Design and Implementation*, 2010, pp. 1–6.

[7] A. Tootoonchian and Y. Ganjali, "HyperFlow: A Distributed Control Plane for OpenFlow," *Proc. Internet Network Management Conference on Research on Enterprise Networking*, 2010, pp. 3–3.

[8] S. Hassas Yeganeh and Y. Ganjali, "Kandoo: A Framework for Efficient and Scalable Offloading of Control Applications," *Proc. 1st workshop on Hot Topics in Software Defined Networks*, 2012, pp. 19–24.

[9] M. R. Nascimento *et al.*, "Virtual Routers as a Service: The Routeflow Approach Leveraging Software-Defined Networks," *Proc. 6th Int'l. Conf. Future Internet Technologies, ser. CFI '11*, New York, NY, USA: ACM, 2011, pp. 34–37.

[10] C. E. Rothenberg *et al.*, "When Open Source Meets Network Control Planes," *IEEE Computer Mag.*, vol. 47, no. 11, Nov. 2014, pp. 46–54.

[11] R. Sherwood *et al.*, "FlowVisor: A Network Virtualization Layer," *OpenFlow Switch Consortium, Tech. Rep.*, 2009.

[12] E. Salvadori *et al.*, "Demonstrating Generalized Virtual Topologies in an OpenFlow Network," *ACM SIGCOMM Computer Commun. Rev.*, vol. 41, no. 4, Aug. 2011, pp. 458–59.

[13] R. Corin *et al.*, "VeRTIGO: Network Virtualization and Beyond," Proc. European Workshop on Software Defined Networking, 2012, pp. 24–29.

[14] O. M. C. Rendon *et al.*, "Monitoring Virtual Nodes using Mashups," *Elsevier Computer Networks*, vol. 64, May 2014, pp. 55–70.

BIOGRAPHIES

JULIANO ARAUJO WICKBOLDT (jwickboldt@inf.ufrgs.br) is a Ph.D. student at the Federal University of Rio Grande do Sul (UFRGS) in Brazil. He achieved his B.Sc. degree in computer science at Pontifical Catholic University of Rio Grande do Sul in 2006. He also holds an M.Sc. degree from UFRGS conducted in a joint project with HP Labs Bristol and Palo Alto. Juliano was an intern at NEC Labs Europe in Heidelberg, Germany for one year between 2011 and 2012. Between 2013 and 2014 Juliano was a substitute professor at UFRGS. His current research interests include cloud resource management and software-defined networking.

WANDERSON PAIM DE JESUS (wpjesus@inf.ufrgs.br) coordinates research and development projects at the National Education and Research Network (RNP) in Brazil. He is also an MBA student in strategic management of information technology at Fundao Getulho Vargas (FGV). Wanderson achieved his M.Sc. degree at the Federal University of Rio Grande do Sul (UFRGS) in 2013, and his B.Sc. at the Federal University of Gois (UFG) in 2010, both in computer science. He is currently committed to foster industry-academia partnerships around information and communication technology. His research interests include software-defined networking and cloud computing.

PEDRO HELENO ISOLANI (phisolani@inf.ufrgs.br) is an M.Sc. student in computer networks at the Federal University of Rio Grande do Sul (UFRGS) in Brazil. He achieved his bachelor's degree in information systems at the State University of Santa Catarina (UDESC) in 2012. During the undergraduate years he worked on scientific initiation research in the management of learning objects development process. Currently, his research interests are related to the management of software-defined networking, more specifically monitoring, visualization, and configuration management activities.

CRISTIANO BONATO BOTH (cbboth@inf.ufrgs.br) is an associate professor at the University of Santa Cruz do Sul, Brazil. He received his Ph.D. degree in computer science from UFRGS in 2011. He received his M.Sc. degree in computer science from the Pontifical Catholic University of Rio Grande do Sul, Brazil, in 2003. His research interests include wireless networks, next generation networks, and traffic control on broadband computer networks.

JUERGEN ROCHOL (juergen@inf.ufrgs.br) is an emeritus professor at the Institute of Informatics of the Federal University of Rio Grande do Sul (UFRGS), Brazil. He received his M.Sc. degree in physics and his Ph.D. degree in computer science, both from UFRGS, in 1972 and 2001, respectively. His research interests include wireless networks, next generation networks, optical networks, and traffic control on broadband computer networks.

LISANDRO ZAMBENEDETTI GRANVILLE (granville@inf.ufrgs.br) is an associate professor at the Institute of Informatics of the Federal University of Rio Grande do Sul (UFRGS), Brazil. He received his M.Sc. and Ph.D. degrees, both in computer science, from UFRGS in 1998 and 2001, respectively. Lisandro has served as a TPC co-chair of IFIP/IEEE DSOM 2007, IFIP/IEEE NOMS 2010, TPC vice-chair of CNSM 2010, and general co-chair of CNSM 2014. He is also chair of the IEEE Communications Society Committee on Network Operations and Management (CNOM), and co-chair of the Network Management Research Group (NMRG) of the Internet Research Task Force (IRTF). His areas of interest include management of virtualization for the Future Internet, Software-Defined Networking (SDN), and Network Functions Virtualization (NFV), and P2P-based services and applications.

A Recommender System Architecture for Predictive Telecom Network Management

Faisal Zaman, Gabriel Hogan, Sven van der Meer, John Keeney, Sebastian Robitzsch, and Gabriel-Miro Muntean

Faisal Zaman, Sebastian Robitzsch, and Gabriel-Miro Muntean are with Dublin City University.

Gabriel Hogan is with the Centre for Global Intelligent Content.

Sven van der Meer and J. Keeney are with Ericsson Ireland.

ABSTRACT

Current telecom networks generate massive amounts of monitoring data consisting of observations on network faults, configuration, accounting, performance, and security. Due to the ever increasing degree of complexity of networks, coupled with specific constraints (legal, regulatory, increasing scale of management in heterogeneous networks), the traditional reactive management approaches are increasingly stretched beyond their capabilities. A new network management paradigm is required that takes a preemptive rather than reactive approach to network management.

This work presents the design and specification of E-Stream, a predictive recommendation-based solution to automated network management. The architecture of E-Stream illustrates the challenges of leveraging vast volumes of management data to identify preemptive corrective actions. Such design challenges are mitigated by the components of E-Stream, which together form a single functional system. The E-Stream approach starts by abstracting trace information to extract sequences of events relevant to interesting incidents in the network. After observing event sequences in incoming event streams, specific appropriate actions are selected, ranked, and recommended to preempt the predicted incidents.

INTRODUCTION

The combined effect of the increase in users and communicating devices, demand for service quality and diversity, support for mobility, and desire for social connectedness and communication has driven unprecedented and exponential growth in telecom network management data. Following this growth of users and devices, by 2020 the total number of connected devices will reach up to 50 billion [1]. Consequently, the number of network elements (NEs) to manage will increase significantly. The amount of network management data transmitted from these NEs is expected to be at exabyte levels. Also, as heterogeneous networks are becoming a reality with the deployment of micro-, femto-, and picocells [1], the complexity of the operation and management (O&M) tasks [2] scales up accordingly. To maintain such complex networks current O&M approaches need to be extended in order to provide efficient and high-quality communication services to end users. This mostly impacts operating costs for operators as today's approaches for monitoring rapidly expanding user and device volumes will require a significant increase in management personnel, which, based on current approaches, is economically unsustainable. Many of the tasks required of human network managers are repetitive and involve wading through huge amounts of monitoring data. A new network management paradigm is required that is capable of automating the monitoring and repetitive tasks, and most importantly leverage massive volumes of network trace information to deploy a preemptive rather than reactive approach to predict issues and suggest timely appropriate remedial or preventative actions for network management.

More automated approaches are required to assist network operations center (NOC) operators to manage complex network operation scenarios. Intelligent techniques with the capability to decipher the *recurrent* nature of predictable network incidents can unravel the link between predictive symptoms and the occurrence of a particular network scenario. Moreover, this insight into patterns of symptomatic events can be leveraged to prescribe potential solution(s) for particular scenarios. Integrating these two functionalities into a single information processing system is proposed in E-Stream. E-Stream analyzes the likelihood of the occurrence of potential network incidents and recommends the most appropriate solution for the incident. In addition to this, E-Stream gives the human operator control to adopt decisions while allowing the system to learn the decision recommendations over time, and adapt and evolve by assimilating response know-how from the human expert.

0163-6804/15/$25.00 © 2015 IEEE

From an operational point of view, E-Stream takes the network traces as inputs and transforms the massive volume of information into simple prescriptive actions as outputs. First, E-Stream discards unnecessary, redundant, and noisy information in order to observe patterns in the occurrence of network incidents. Patterns are then screened, indexed, and associated with the relevant network solutions. Finally, the "pattern-solution" templates of similar incidents are utilized to suggest proper corrective actions for network scenarios similar to those in which similar patterns were previously observed. The performance of the components at each phase is scaled by ingesting the incoming traces in an adaptive way, distributing the processing resources, and parallelizing the computational tasks.

Autonomic network management (ANM) [3] emerged as an approach to overcome the ever increasing complexity of network management within the fault, configuration, accounting, performance, security (FCAPS) framework. Efficient fault management (FM) techniques, proposed as part of an autonomic network management system (ANMS), should be able to progressively learn and identify network faults but do not present solutions [4]. A similar approach [5] for preemptive detection of critical events in the area of service management proposed a process to discover potential predictive patterns in the log files to detect the occurrence of upcoming faults. Network management based on preventative maintenance and statistical process control (widely used in manufacturing industries) was proposed in [6]. However, the E-Stream approach for predictions and recommendations is innovative in two main aspects. First, it supports processing events from heterogeneous sources, and in this way overcomes classic problems of FCAPS silos. Second, E-Stream addresses the requirements of the telecommunication management network (TMN) in terms of scale and performance. E-Stream combines both aspects into a holistic system design and implementation that addresses some of the challenges the telecommunication industry is facing today. Additionally, the process of providing corrective actions based on the predicted network incidents, and with minimal human supervision, is novel and very different from the approach of gradually learning about the network faults in an ANMS.

This work is the result of a collaborative Dublin City University-LM Ericsson Ireland project that integrates results of data mining, predictive analytics, and recommender systems research.

E-Stream Challenges

The major challenge in designing E-Stream is the complex granular structure inherent in network data; such complex structures in the data complicate efforts to automatically find meaningful information from the data. The task is to transform the incoming information through different processing layers and finally deliver the recommendations. It requires expert knowledge on the deployment and management of the resources in each layer. To do so, E-Stream is composed of several components, and each component addresses and mitigates the challenges in processing the information and transferring it to the next layer. In Fig. 1 the components of the E-stream architecture are placed in appropriate processing layers. The challenges addressed by each component in each layer are also indicated.

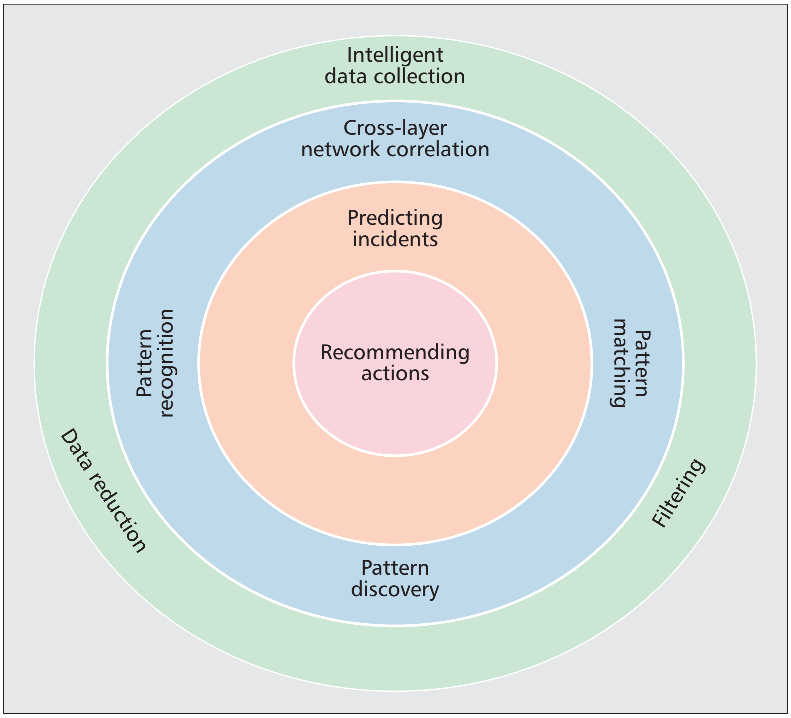


Figure 1. E-Stream components in different information processing layers.

Intelligent Data Collection

A stream of network traces (denoted as e-streams) originate from the telecom networks often arrive at extremely high rates, straining the I/O and computational ability of the system. The traces consist of information patterns that can be correlated to the underlying network behavior. E-streams are generated from various sources, and report on a huge variety of parameters, operations, warnings, and faults, only some of which are relevant for any particular management use case. Combined with a potential for event storms, this in turn makes it very difficult to extract meaningful knowledge from the data within a limited time duration.

Data dimension reduction processes aim to reduce the volume of data being ingested by identifying and removing noise events, low importance events, and periodic or repetitive events. Dimension reduction processes therefore aim to identify the most important events, and data within those events. Simple data reduction techniques can serve the purpose of controlling data ingestion, but it should be capable of handling sizeable volumes of data. Parallelizing the reduction process can address this problem, but raises the possibility of more approximation errors. Event-based stream processing (ESP) holds the most promise for parallelized event reduction processes with low latency, but care must be taken to maintain scalability [7].

The system therefore needs to be able to differentiate between audiences and facilitate recommendations based on their individual time/response characteristics, the degree of competence of the individual user, and the level of autonomy of the system.

Cross-Layer Network Correlation

Network traces (e-streams) are not static and demonstrate burstiness, jitter, delay (out-of-order arrival), and data loss. In order to explore e-streams, the underlying individual sources need to be analyzed to discover, recognize, and match patterns across all the sources. These patterns can be indicative of correlative scenarios that are difficult to decipher from individual sources. Pattern discovery is required to detect the existence of patterns that have not previously been observed. Pattern matching in the event stream is required to indicate the occurrence of previously observed patterns. Pattern recognition is required in order to determine the likelihood of a candidate pattern becoming an exact match, and to allow the prediction of and therefore the prevention of an incident occurrence.

For event-based pattern matching, complex event processing (CEP) is a recognized technique. Topologically-aware reasoning (TAR) addresses the problem of discovering and matching patterns to identify network faults based on spatio-temporal patterns [8]. Automated profiling of network events by modeling the event sequences is beneficial for pattern matching [9].

Predicting Incidents

Predicting incidents is dependent on pattern recognition accuracy, that is, the probability of correctly identifying a complete pattern from first symptom to incidence occurrence. Incidents can be categorized at a high level with the following characteristics: incidents occur frequently or rarely in time; incidents are simple or complex; incidents have a simple or complex resolution. System behavior (event patterns) characterizing an incident can present as a small number of symptoms in a single data set or as a very large number of symptoms across many data sets. Predicting incidents therefore ranges from those that occur very often and have very simple analysis to those that occur rarely and involve very complex analysis. Temporal analysis is constrained by the amount of observation that can be supported in any given time period, that is, the available time for observation is inversely proportional to the volume of data being observed. The recognition accuracy is constrained by the number of candidate (probable) patterns being observed.

Recommending Actions

Action recommendation is dependent on a number of factors including context, audience, existing action/responses, and validation.

Context — In a typical telecommunication system when an incident occurs, one of a number of formal and prescribed responses to the incident is normally followed. This typically has instructional and procedural aspects, that is, the specific tasks to be carried out, the order or priority of the action/response, and the reporting and tracking of the incident through, for instance, trouble ticketing systems.

Audience — A number of possible audiences with different time/response characteristics interact with the network management system: human operators with different roles, authorities, and competence levels; various integrated response systems such as trouble ticketing systems; and workflow and process management systems. Each of these audiences require different actions suited to their view of the system and its behavior. The system therefore needs to be able to differentiate between audiences and facilitate recommendations based on their individual time/response characteristics, the degree of competence of the individual user, and the level of autonomy of the system.

Existing Responses — The majority of responses are based on previously performed best practice responses; for example, many faults have detailed specific instructions that are followed to resolve the issue. However, it is important to realize that best practice responses differ for different networks, operators, and customers. For established networks, this knowledge covers a large percentage of known faults and forms a body of pre-existing responses to known incident types. However, this body of knowledge of reactive actions may not be effective for preemptively preventing or dealing with incidents before they occur. In addition, where a specific response does not exist or an incident occurs for the first small number of times, this can be characterized as a *cold start* problem.

Validation — When recommendations are suggested, they have to be sanity checked by the domain expert, and the accuracy of the recommended actions has to be validated. The accuracy of validation has to be a learning component for future recommendations; that is, the recommender has to weigh the correctness of previous recommendations, and adjust the production and ranking of future recommendations accordingly. After initial training with the domain expert, the recommender systems must learn and provide recommendations to a level such that the operator reduces the degree of supervision (eventually close to zero).

E-Stream Components

E-stream is a composite system to recommend corrective solutions to network scenarios based on predictive patterns leading up to network incidences. These predictions and recommendations are built up from several independent components interacting in stages. Each component is capable of carrying out different functionalities:

- Data reducer: reducing the incoming data
- Correlator: filtering out irrelevant events and correlating events
- Pattern matcher: modeling and matching patterns
- Predictor: predicting future patterns
- Recommender: recommending solutions

Data Reducer and Correlator

The data reducer and correlator component builds up a *smart reduction* mechanism to extract *actionable insights* from network traces that are utilized by other components in the later stages. As discussed, the bursty nature and inherent variability in the trace sources further complicate the trace exploration. Data reducer mitigated the challenge of data-deluge through *controlled ingestion* of event traces into the I/O system and deploying *scaled projection* to compress the event trace information. The correlator examines the significance of events analyzing the inter-event relationship through spectral distribution and temporal dependence, and envelops the relevant event traces into pseudo patterns. Accurately removing irrelevant and insignificant events from the raw data stream increases the probability of accurately finding coherent event types. Also, running sequence mining techniques on reduced events is computationally much cheaper than searching for event associativity in raw data.

The functional architectures of the data reducer and correlator are entwined as shown in Fig. 2. The data reducer is equipped with window based *dynamic* load shedding and Johnson-Lindenstrauss Theorem (JLT) [10] based minimal loss approximation functionality. The correlator operates with frequency and spectral domain-based filters and sequence mining techniques.

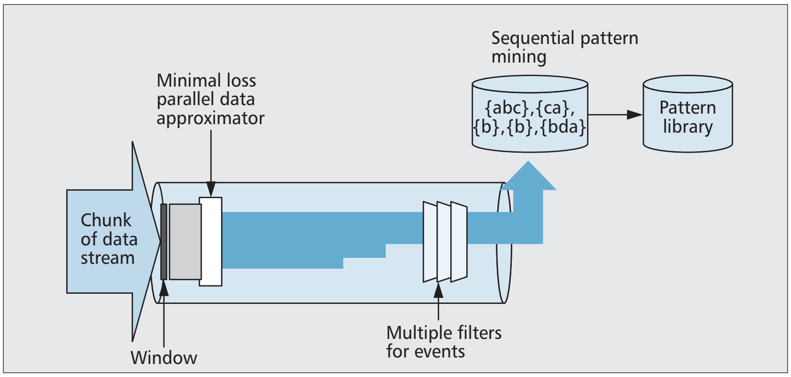

Sequential pattern mining
Minimal loss parallel data approximator
{abc},{ca},{b},{b},{bda}
Pattern library
Chunk of data stream
Multiple filters for events
Window

Figure 2. The data reducer consists of the window and minimal loss parallel data approximation; the correlator is composed of multiple filters and sequence miner.

Windowing — The concept of dynamic windowing is based on *controlled* ingestion of data streams. This entails feeding the processors with manageable volumes of data while dealing with sudden and extreme influxes of event traces. The objective of the dynamic windowing process is to maintain a maximum end-to-end latency of the overall system. In this mechanism the volume of incoming data is adaptively controlled based on the data arrival rate (also defined as stream burst rate), that is, *automatically* change the data-read rate (length of the window) based on the stream burst rate. Leveraging the data arrival rate distribution allows E-Stream to control the volume of incoming data in line with the capacity of the processors (buffer size).

Minimal Loss Parallel Data Approximation — To mitigate against very high dimensional bursty data streams, a computationally cheap dimensionality reduction technique, minimal loss parallel data approximation (MLDA), is devised. MLDA reduces event traces while efficiently approximating the degree of correlation (distance) between the events.

The basic principle of data approximation in MLDA is based on a JLT approach. In simple terms, JLT operates according to the principle that if data points of a high dimension vector space are projected into a randomly selected subspace of sufficiently low dimensionality, with high probability the proportional distances between pairs of data points are preserved with a certain level of approximation. Principal component analysis (PCA), another state-of-the-art statistical reduction technique, is more accurate in approximating the data and reducing dimensions, but PCA is computationally infeasible for very high-dimensional data [11]. In order to reduce the distortion of JLT, a minimal loss approximation is proposed. The minimal loss criterion is based on an extension of JLT by combining a Chernoff bound: if the projection is repeated $O(\log(1/\delta))$ times, and the median of the distance between the projections is taken, the probability of accurate approximation is increased to $1 - \delta$. This procedure can be represented in terms of the Law of Large Numbers (LLN), whereby performing an estimation procedure function multiple times (e.g., repeatedly measuring distance between sampled and original points) will increase the probability of accurate estimation of the function.

In summary, MLDA embeds N events recorded in L timeslots of length Δ_t into a space of lower dimension M, such that all distances are almost preserved through the transformation operator Φ. The scheme is shown in Fig. 3, where matrix A represents the event. Iterating this procedure of approximating the distances multiple times following the principle of LLN can produce a more condensed stream of event traces over which the correlation techniques can find pseudo patterns of events more accurately [12].

Online Filtering — The functionality of event filtering is akin to data reduction with the objectives to abstract low-level event information in a more structured way and to quantify the interrelationships between network event traces. The events are reduced by finding correlated events, defining cluster event prototypes, and filtering out noise events. Isolated network events such as periodic reporting events or routine configuration events appear in a standalone manner and are unrelated to other network events, and thus show low correlations with other events around them; therefore, they can be identified and treated as noise events.

Based on this objective, the following online filters are deployed:

- Spectral filter to find correlations and remove noise
- Temporal filter to define cluster of events based on the temporal distances of co-occurrence
- Periodicity filter to filter out periodic events

The principle of the spectral filtering technique is based on random matrix theory (RMT). According to RMT, a confidence band derived from the eigen-value distribution of random matrices can be utilized to separate the true signal from the random noise of a correlation matrix. The spectral filter analyzes the eigenspace (spectrum) of the correlation matrix of the

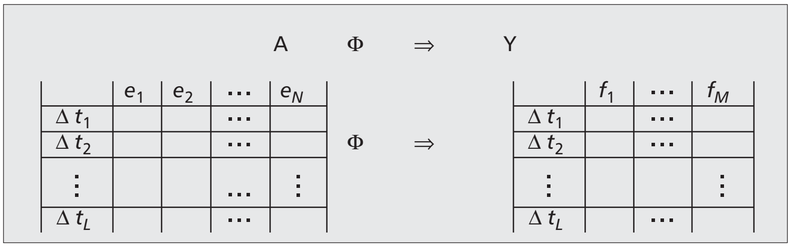


Figure 3. Data reduction through transformation operator Φ.

observed events and decompose the matrix into two parts, one part exhibiting strong correlative structure between the events and another part with weak spectral condition. The later part is treated as noise and is removed from the stream by the spectral filter. In this way the spectral filter acts as both correlation and noise filter.

The temporal distance between symptoms and effects of any network incident, even spread over several different windows, are approximately homogeneous. Therefore, statistically, events appearing together within a *specific time distance* can be defined as a cluster of contiguous events. The temporal filter applies temporal distance-based clustering to find these clusters in the correlated events.

In order to extend noise filtering with the ability to detect event patterns with periodic occurrences, a filter based on frequency domain is embedded after the temporal filtering. After noise is removed and the data density increased, the final task is to define the sequential relationships between the correlated events.

Sequence Miner — Correlated event traces reveal only superficial relationships between the events. In order to extract patterns from the correlated events, sequential mining techniques are applied. Association rule mining algorithms are capable of exploring the sequential relationship between the events to identify the order of occurrence and degree of association between events. This component employs association rule mining techniques over the correlated event traces to identify how event sequences are associated with actual incidents and forms an event pattern. The magnitude of the association is quantified by several metrics. Once association rules between the events and incidents are established, these pattern rules are stored in a pattern library.

Pattern Matcher and Predictor

This component carries out two main tasks:

- Encode and model patterns based on the association probability metrics drawn from the sequence miner.
- Predict the occurrence of event patterns based on matching the occurrence of some of the events (the pattern "head") in the pattern.

Pattern Modeling — This component uses the pattern definitions (association rules) and their associativity metrics from the sequence miner to formulate pattern models. First, statistical similarity analysis is used to discover the relevance between the event sequences in the pattern definitions. The similarity between the event sequence(s) preceding a response (result or consequence) sequence provides the semantics to define a pattern model. This information about the relationships between antecedent-response sequences is important later for the recommender to assess the accuracy of the recommended actions for the predicted incidents. Hamming distance-based locally sensitive hashing (LSH) is used here to compute the similarity between the pattern models.

Conventional associativity metrics (*support*, *confidence*, and *lift*) are computed for each event pattern to characterize the formulated pattern model. The support counts the frequency of an event pattern, confidence computes the probability of a specific antecedent-response forming an event pattern, lift calculates the likelihood of co-occurrence of a specific antecedent-response pair. The knowledge provided by the metrics can be leveraged to determine the probability of a certain event sequence and incident forming a pattern model, and hence accurately predict that incident whenever the event sequence is observed.

Pattern Recognition and Matching — Pattern definitions or rules of association between the set of event sequences of a pattern model are exported into Extensible Markup Language (XML) using Predictive Model Markup Language (PMML). These pattern definitions are then matched in the incoming stream using an off-the-shelf CEP engine, Esper.[1] When an exact match occurs (i.e., 100 percent probability that an incident has occurred), a notification is sent to the recommender with the details of the incident.

Prediction — The metric values along with the antecedent-response pairs characterize each pattern model, and these are used as attributes of the models. Running a supervised learning paradigm over these attributes of the pattern models can identify appropriate "tail" event sequences for a given sequence head of a pattern model [13]. As each event in a pattern is observed in the pattern models, the supervised models should be able to predict the pattern tails with increasing probability.

From Predictions to Recommendations

When an incident occurs, or its pattern tail/response is predicted, all relevant information is forwarded to the Recommender. The Action Selector then selects a set of candidate corrective actions from the action catalog, informed by relevant and similar antecedent-response pairs and associated human-informed actions that were previously applied. The action list is then sorted and ranked by the action ranker based on ranking parameters derived from guidelines of best practice and/or the historical adoption of similar actions in response to similar past incident indications. A ranked list of suggested actions may then be further manipulated, for example, by selecting only the highest ranked suggestions. While more than one recommendation is presented, the recommenda-

[1] EsperTech: http://esper.codehaus.org

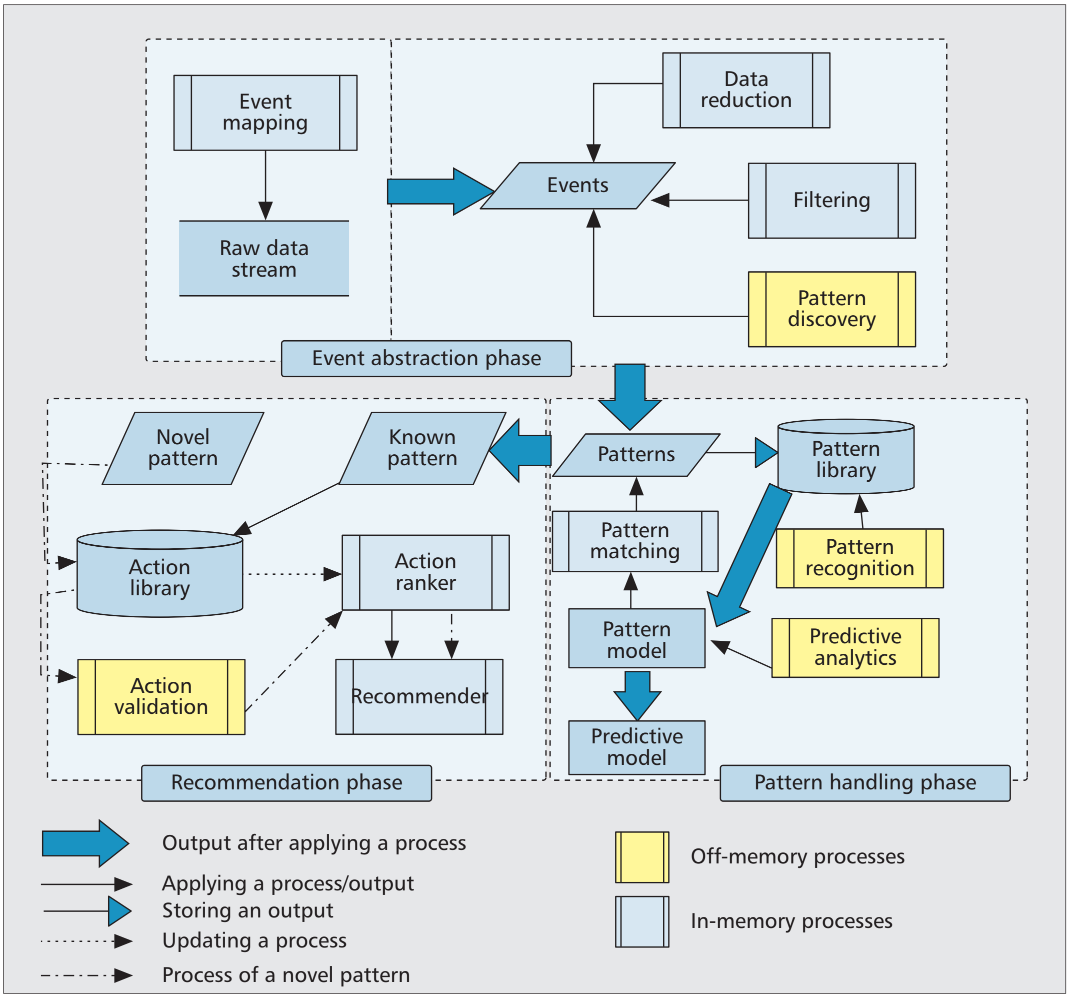


Figure 4. Different transformation phases and computing platforms for processing the network trace information.

> *This feedback mechanism provides a way to deal with the "cold start" problem inherent in all recommender systems. This approach also provides a mechanism to evolve as the network, context, institutional knowledge, or business priorities for the network evolves.*

tions are ranked, so the management agent can quickly see alternative recommendations to a given incident or incident prediction, and the degree to which those recommendations are deemed suitable. The manager then retains the authority to re-rank the list, select one or more of the recommendations, ignore all recommendations, or explicitly select, define, or refine a different action. This step provides feedback to the recommendation system to adjust or extend its action catalog and refine its ranking parameters. Using this continuous feedback process, the recommender system learns and evolves. This feedback mechanism provides a way to deal with the "cold start" problem inherent in all recommender systems. This approach also provides a mechanism to evolve as the network, context, institutional knowledge, and business priorities for the network evolve.

Information Processing

E-stream components perform online processing of incoming information (the e-streams), transfer outputs to the next component in turn for further processing, and finally recommend solutions to various network incidents and incident predictions. Each processing step is designed to run on appropriate computing platforms to maintain scalability of E-stream as an end-to-end system. The flow and transformation of the traces is depicted in Fig. 4. In the figure, the processes running on different computing platforms are indicated with separate colors.

Information Flow

Streamed trace information flows through each component and undergoes different transformations to finally trigger an action (or a set of actions). Based on the objective and functionality of the components, the transformation of the traces can be categorized into three phases.

Event Abstraction Phase — In this phase high-volume event traces (e-streams) originating from different network management sources are compressed into an abstract form defined as patterns. At each step of this phase, the volume of data is reduced and abstracted by a data reducer, filters, and sequential miner components of E-Stream. Event information is extracted from the input raw data using event mapping. Events are matched with the respective sources and normalized. The data reduction procedure then syncronizes the events incurring minimal data information loss. Filtering techniques then correlate events, remove noise, and cluster similar event types. The ordering of the filtering tasks is based on the context of the trace; removing noise increases the probability of accurately finding relevant event types. The final outcomes of this phase are the discovered event sequences incorporating association rules and the event relationship metrics.

Pattern Handling Phase — In this phase, patterns stored in the pattern library are used to recognize patterns and build *pattern models*. The patterns are then matched in the incoming streams. The association rules and relationship metrics from the pattern models are then used to calculate the increasing likelihood of earlier events in a pattern being able to predict later events in each particular pattern. In the supervised learning framework, an incident is predicted whenever the probability is higher than a predefined threshold of certainty.

The sequential pattern discovery task is specifically implemented to utilize disk based processing for high throughput. Due to the data intensive nature of pattern recognition tasks, pattern model building for matching patterns, and prediction of patterns, these task are required run off-memory.

Recommendation Phase — In the third and last phase, *actions* are suggested for *known* patterns, which were successfully recognized, matched and predicted in the previous phase. For matched patterns, specific actions are recommended from the action library; for the predicted patterns, the actions from the library are first ranked based on the action-response relationship, and then the top actions are recommended. Event sequences without any prior profile (or unrecognized patterns) are defined as *novel* patterns. Actions for these patterns are first selected based on the proximity with other pattern models in the pattern library. Then these set of actions are validated in order to be able to recommend the most appropriate actions.

Scalability and Optimization of Information Processing

Online information processing has three limiting factors: scale, precision, and timeliness. To address scale, components are designed to exploit parallel computation and distributed processing tools. We consider two different aspects of parallel computing here, pipeline parallelism and partial parallelism, to execute the tasks of each component. Based on the throughput, tasks are implemented on in-memory and off-memory processing platforms to facilitate restricted latencies. In Fig. 4, the processes running in-memory are colored light blue and the processes running off memory are colored yellow.

In-Memory Processing — In E-stream, data reduction, filtering, pattern matching, and action recommendation for matched patterns require the data structure to be preserved, for information to be passed quickly and easily through access points, and finally, need to process the information in near real time. In-memory processing is most suitable for these complex and time-sensitive tasks to achieve increasing speed and reliability to deliver the output within a limited time delay.

Off-Memory Processing — Off-memory processing is typically disk-based, meaning the application queries data stored on off-RAM. In contrast to in-memory, off-memory processing can deal with huge amounts of data. The sequential pattern discovery task is specifically implemented to utilize disk-based processing for high throughput. Due to the data-intensive nature of pattern recognition tasks, pattern model building for matching patterns, and prediction of patterns, these task are required to run off-memory.

Among the analytical tasks (colored in yellow in Fig. 4), data reduction and filtering are realized using the Storm[2] stream computing framework. Pattern matching is realized using the Hadoop-MapReduce framework, which leverages the MapReduce computing framework for the exact matching tasks.

Conclusion

Large complex systems such as telecommunications networks are very difficult and costly to manage. Predicting, preventing, mitigating, and fixing problems as early in the discovery cycle as possible is a key strategy in reducing operational expenditure. The Predict & Recommend approach presented in this article has the advantage that the network management system benehuman experts' approaches. Human operators should retain the ultimate decision to ignore or select from the recommendations presented until the system provides accurate and confident predictions with high-value benefits. Thus, the recommender system starts as an assistant, but can later have authority to perform automated tasks delegated or revoked. This allows human operators to concentrate on high-value cases, exceptions, and critical situations not yet sufficiently learned by the prediction and recommender systems.

Using prediction and recommender systems as presented in this article will support faster resolution of network ssues by presenting candidate solutions, rather than simply presenting a list of incidents. As the system learns, tuning best practice for a given network deployment and behavior characteristics, the amount of mundane troubleshooting required by NOC personnel for day-to-day operation and maintenance of the network is reduced, thus reducing cost, improving management throughput, and freeing up time and resources for the managers to concentrate on more strategic management issues.

References

[1] Ericsson, "More than 50 Billion Connected Devices," http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf; Ericsson Whitepaper: Feb. 2011.

[2] Infonetics Research, *Subscriber Data Management Software and Services*, 2nd ed., Nov. 2011.

[3] B. Jennings *et al.*, "Toward Autonomic Management of Communications Networks," *IEEE Commun. Mag.*, vol 45, no 10, 2007, pp. 112–21.

[4] R. Chaparadza, "UniFAFF: A Unified Framework for Implementing Autonomic Fault Management and Failure Detection for Self-Managing Networks," *Int'l. J. Network Mgmt.*, vol. 19, no. 4, 2009, pp. 271–90.

[5] W. Zirkel and G. Wirtz, "A Process for Identifying Predictive Correlation Patterns in Service Management Systems," *Proc. Int'l. Conf. Service Sys. and Service Mgmt. '10*, 2010, pp. 1–6.

[6] J. Keeney, S. van der Meer, and G. Hogan, "A Recommender-System for Telecommunicatons Network Management Actions," *Proc. IFIP/IEEE. Symp. on Network Management 2013—TechSessions,* Ghent, Belgium, 2013, pp. 760–63.

[2] http://storm-project.net/

[7] A. Brito *et al.*, "Scalable and Low-Latency Data Processing with Stream MapReduce," *Proc. IEEE CloudCom*, 2011, pp. 48–58.

[8] T. Wang *et al*., "Spatio-Temporal Patterns in Network Events," *Proc. ACM Int'l. Co-NEXT '10*, article 3, 2010.

[9] X. Meng *et al.*, "Automatic Profiling of Network Event Sequences: Algorithm and Applications," *Proc. IEEE INFOCOM '08*, 2008, pp. 266–70.

[10] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz Mappings into a Hilbert Space," *Proc. Conf. Modern Anal. & Probability*, New Haven, CN, 1984, *Contemporary Mathematics 26*, 1984, pp. 189–206.

[11] D. Fradkin and D. Madigan, "Experiments with Random Projections for Machine Learning," *Proc. 9th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining,*, ACM, pp. 517–22.

[12] G. Reeves *et al*., "Managing Massive Time Series Streams with Multi-Scale Compressed Trickles," *Proc. VLDB Endow*., vol. 2, no. 1, Aug. 2009, pp. 97–108.

[13] C. Rudin, B. Letham, and D. Madigan, "Learning Theory Analysis for Association Rules and Sequential Event Prediction," *J. Mach. Learning Research*, vol. 14, no. 1 Jan. 2013, 2013, pp. 3441–92.

Additional Reading

[1] F. Zaman *et al*., "A Heuristic Correlation Algorithm for Data Reduction Through Noise Detection In Stream-Based Communication Management Systems," *Proc. IEEE NOMS*, pp. 1–8.

[2] E-Stream: Innovative Performance Management System for Heterogeneous Networks, http://www.estream-project.com.

Biographies

Faisal Zaman (zaman.faisal@ieee.org)is a postdoctoral researcher with the Performance Engineering Laboratory and Network Innovations Centre, Rince Institute, Dublin City University, Ireland. He received his Ph.D. in information science from Kyushu Institute of Technology in 2011. In his previous tenure as a postdoctoral researcher at Kyushu Institute of Technology, he analyzed time series data for weather forecasting and micro-array data for gene classification. He also worked as a statistical programmer in Shafi Consultancy Ltd and led analytical teams to analyze medical trial data. He is a program committee member of several data mining conferences. He has published 30 articles, conference proceedings, books, book chapters, conference papers, and technical reports. He has experience in supervising Ph.D. and M.Sc. level students.

Gabriel Hogan (ghogan@cngl.ie) is the intellectual property manager at the Centre for Global Intelligent Content (CNGL). His primary focus is enabling research impact through collaborative innovation and entrepreneurship. He holds a B.Sc. in software and mathematics, and an M.Sc. in technology and innovation management, and has over 25 years' experience in ICT in the corporate, startup, and academic sectors. He was previously head of research and innovation at LM Ericsson Ireland where he directed a number of research teams and chaired Ericssons' Network and Service Management Patent Board. He has published on and filed intellectual property across the areas of efficient mobile power, recommender systems, predictive analytics, stream processing, and network performance analysis. He has coordinated and led a number of industry/business centred EU and Irish government funded research projects.

Sven van der Meer (vdmeer@ieee.com) received his Ph.D. in 2002 from Technical University Berlin. He joined Ericsson in 2011, where he is currently a master engineer leading a team that will enhance the capabilities of Ericsson's OSS products. Most of his current time is dedicated to designing and building advanced policy and predictive analytics systems. In the past, he has worked with Fraunhofer FOKUS (Berlin, Germany), Technical University Berlin (Germany) and the Telecommunication Software and Systems Group (TSSG, Ireland), leading teams and projects, consulting partners and customers, and teaching at the university level. He is actively involved in the IEEE CNOM community as a standing member of program committees (IM, NOMS, CNMS, and APNOMS among others), and has helped to create and organize successful workshop series (MACE, MUCS, and ManFed.Com, among others). He has also contributed to the OMG and TM Forum standardization organizations. He has published in more than 100 articles, conference proceedings, books, book chapters, conference papers, and technical reports. He has supervised and evaluated 6 Ph.D. and more than 30 M.Sc. students.

John Keeney is a senior researcher in LM Ericsson Ireland, working in the Network Management laboratory in Ericsson's Software Research Campus in Athlone. His research focus is on monitoring and managing complex systems, especially telecoms systems, with a particular focus on knowledge extraction, event stream processing, and performance analysis. His work in Ericsson centers on online analytics and optimization of radio access network performance to inform the next generation of operation system support concepts for Ericsson's OSS product unit.

Sebastian Robitzsch is a postdoctoral researcher with the Rince Institute at Dublin City University. He received his Ph.D. from University College Dublind in 2013 and an M.Sc. equivalent (Dipl.-Ing. (FH)) from the University of Applied Sciences Merseburg, Germany. In the past he has worked with T-Systems, Germany; Fraunhofer FOKUS, Germany; and Nokia Research Centre, Finland. His research spans from interference issues and self-configuration techniques in 802.11-based multi- antenna mesh networks, heterogeneous radio access networks to system architecture design for trace analytics and recommender systems for next-generation OSSs. Furthermore, he is involved in ongoing European research activities focusing on the fifth generation of mobile networks.

Gabriel-Miro-Muntean [M] (gabriel.muntean@dcu.ie) received his Ph.D. degree from Dublin City University for research in the area of quality-oriented adaptive multimedia streaming in 2003. He is a senior lecturer with the School of Electronic Engineering at Dublin City University (DCU). He is a co-director of the DCU Performance Engineering Laboratory and consultant professor with Beijing University of Posts and Telecommunications, China. His research interests include quality-oriented and performance-related issues of adaptive multimedia delivery, performance of wired and wireless communications, energy-aware networking, and personalized e-learning. He has published over 180 papers in prestigious international journals and conferences, has authored three books and 15 book chapters, and has edited six other books. He is an Associate Editor of *IEEE Transactions on Broadcasting*, Associate Editor of *IEEE Communications Surveys and Tutorials*, and a reviewer for other important international journals, conferences, and funding agencies. He is a member of ACM and the IEEE Broadcast Technology Society.

Using prediction and recommender systems as presented in this article will support faster resolution of network issues by presenting candidate solutions, rather than simply presenting a list of incidents.

Management Architecture for Location-Aware Self-Organizing LTE/LTE-A Small Cell Networks

Sergio Fortes, Alejandro Aguilar-García, Raquel Barco, Felix Barba Barba, Jose Antonio Fernández-Luque, and Alfonso Fernández-Durán

Sergio Fortes, Alejandro Aguilar-García, and Raquel Barco are with Universidad de Málaga.

Felix Barba Barba, Jose Antonio Fernández-Luque, and Alfonso Fernández-Durán are with Alcatel-Lucent.

ABSTRACT

This article proposes a novel architecture for next-generation cellular networks in indoor scenarios. The objective of this model is to support SON mechanisms based on the knowledge of the user equipment location in a small cell network, the use of such information being a key enabler for advanced SON methods. The defined design is the basis for systems providing innovative location-aware SON techniques that make use of user localization in medium/large indoor areas (e.g., malls or corporate buildings). The functional and physical characteristics of this architecture and their technical implications are analyzed. Proposed innovations to generic mobile architecture are described as well as specific implementation for LTE/LTE-A standards. Interoperability with standard management systems and localization services, congestion avoidance, and data offloading are the key drivers of the design. Finally, the capabilities of the proposed architecture are demonstrated through the performance analysis of a simple key use case for location-aware self-optimization.

INTRODUCTION

In recent years, the mobile market has seen the arrival and deep penetration of smartphones and tablets, and an increase in data traffic due to flat rate plans as the main commercial provision scheme for these devices. This situation leads to continuous traffic growth for the mobile network at reduced revenues per user [1]. This is pushing service providers to improve network performance and reduce operational expenditures (OPEX).

In this field, the *self-organizing network* (SON) is a new paradigm defined under the auspices of the Third Generation Partnership Project (3GPP) and the Next Generation Mobile Networks (NGMN) Alliance aiming to automate mobile infrastructure operation, administration, and management (OAM). Via automation and continuous tuning and maintenance of the infrastructure, operators seek to achieve optimum performance at minimum costs.

Here, 3GPP has developed a standardization effort to define the different use cases and architecture integration of the SON functions [2], which act in three main areas:

- *Self-configuration*: the plug and play capabilities of network elements
- *Self-optimization*: the adjustment of parameters during the operational life of the system
- *Self-healing*: failure detection, diagnosis, compensation, and recovery of the network

Additionally, one of the key challenges for the mobile market is to find killer applications for the new terminals and data plans in order to increase service providers', manufacturers', and application developers' revenues. In this field, *location-based services* (LBS) will support key new applications in several fields: advertising, domotics, health care, emergency response, and so on. The most suitable scenarios for LBS are medium/large indoor areas (e.g., corporate buildings, malls, hospitals, airports) given the large concentration of users in these environments. However, these indoor areas impose very challenging conditions for the provision of mobile connectivity: moving obstacles, variable traffic distribution and user movement, infrastructure proneness to failure/outage, and so on.

Therefore, providing coverage for indoor scenarios is becoming a huge challenge due to the high cost of infrastructure. This is especially dramatic for the provision of new mobile communication standards working at high frequencies (e.g., Long Term Evolution, LTE, reaching 2.6 GHz in Europe), where macrocell coverage will be greatly reduced indoors. Operators seek to solve this issue by deploying small cells (low-powered radio access nodes). Different kinds of small cells have been especially defined for indoor scenarios: *picocells*, with up to 200 m coverage; and *femtocells*, with coverage in the range of tens of meters and connected to the mobile network operator via

0163-6804/15/$25.00 © 2015 IEEE

broadband Internet connection (e.g., digital subscriber line, DSL) [3]. In these scenarios, information about the position of user equipments (UEs) can be key for SON algorithms to efficiently configure, optimize, and maintain the network.

In this article, we propose a novel OAM architecture where SON systems could benefit from the increasing number of deployments for indoor positioning to enhance network performance as well as users' satisfaction. In the literature, the position of mobile terminals has scarcely been investigated to support SON mechanisms in macrocell scenarios. Additionally, mechanisms proposed in macrocell references are restricted to a narrow field of network performance. For instance, for monitoring, [4] presents an approach to generate RF coverage maps from the information received from user terminals, the positions of which are estimated by localization techniques. Moreover, [5] develops a fast handover algorithm based on positioning for macrocell scenarios.

Additionally, in order to allow the practical development of location-based SON techniques under the particular conditions of indoor scenarios (coverage overlapping, variable user distributions, cell load, etc.), the support of a complete OAM architecture able to integrate available location services with SON mechanisms is deemed indispensable. This is a novel approach where little work on the concept has been developed in the literature. In outdoor SON schemes, often user traces are roughly localized using cellular-based localization systems in order to provide a geographical perspective on the network status. An architecture model for cellular networks based on a mobile positioning system to assist load balancing methods is briefly depicted in [6], but no details about the interfaces or the OAM module are specified. Also, the considered cellular localization techniques and approaches are not applicable in indoor environments due to the complexity and variability of indoor conditions and propagation delays. Recent projects such as BeFEMTO's present an architecture that includes the use of *local location* [7] as a SON enabler but does not integrate it with end-user localization services; nor does it analyze the impact or the possibilities of location-aware self-organizing mechanisms.

In conclusion, a specific small cell architecture able to integrate external location services is deemed necessary for supporting location-based SON in indoor scenarios, being the objective addressed in this work.

This article is organized as follows. The following section analyzes the particularities of the use of position information for SON mechanisms. The third section defines the proposed OAM architecture. In addition, a model for the specific case of the LTE/LTE-Advanced (LTE-A) standard is presented. In the fourth section a specific algorithm for architecture evaluation is described. Quantitative justification of this architecture and its evaluation is obtained via the system simulator presented in the fifth section. The final section presents the conclusions of this work.

Location Awareness Contribution to SON Mechanisms

Recently developed mechanisms for indoor positioning are based on different technologies: WiFi, mobile phone signal, radio frequency identification (RFID), near-field communication (NFC), inertial systems, and so on. Nowadays, most of these techniques are implemented in mobile phone devices, such as smartphones and tablets.

In this way, precise information about the position of UEs is becoming a common asset for smartphone applications; consequently, industry is starting to take advantage of it in order to offer a wide diversity of location-based services.

The integration of UE localization and network information would allow location-based SON algorithms applicable to multiple network management use cases [8]: coverage and capacity optimization (optimizing coverage to real-time user distributions), energy saving (minimizing consumption where no users are close to a site), load balancing and mobility robustness optimization (finely tuned to current user positions), detection and diagnosis of failures (based on device locations), and so on. For instance, network performance can be improved greatly by dynamic adaptation of small cell coverage areas based on user distribution, reducing interference and power consumption.

Coordination and proper trade-off between different SON use cases is currently a hot topic in literature. Such techniques can also benefit from an integrated architecture able to provide the same location information to different SON mechanisms, avoiding possible collisions due to the use of heterogeneous data.

Coordination and proper trade-off between different SON use cases is currently a hot topic in literature. Such techniques can also benefit for an integrated architecture able to provide the same location information to different SON mechanisms, avoiding possible collisions due to the use of heterogeneous data.

However, one of the main challenges for comprehensive usage of indoor localization in SON mechanisms consists of the definition of an architecture able to provide the positioning information to the OAM plane of the network operator, where positioning and SON have very different characteristics and restrictions.

On one hand, indoor location implementations normally follow a UE- centered approach, where each terminal communicates with a localization manager/server able to perform the (computationally expensive) positioning algorithms and store the (memory-heavy) indoor maps. Also, as users are continuously moving and changing positions, the timescales considered are very small, on the order of seconds or milliseconds.

On the other hand, traditional OAM functions have time spans of hours or days (e.g., the change of small cell parameters is typically performed once a day). SON functionalities, if present, are deployed as part of the OAM architecture. For future networks, a reduction in SON time spans to minutes is foreseen [9]. However, in comparison, localization is still much more dynamic in nature.

Therefore, an architecture that supports the use of localization information by SON functions should cope with these heterogeneous needs by making SON functionalities faster and more

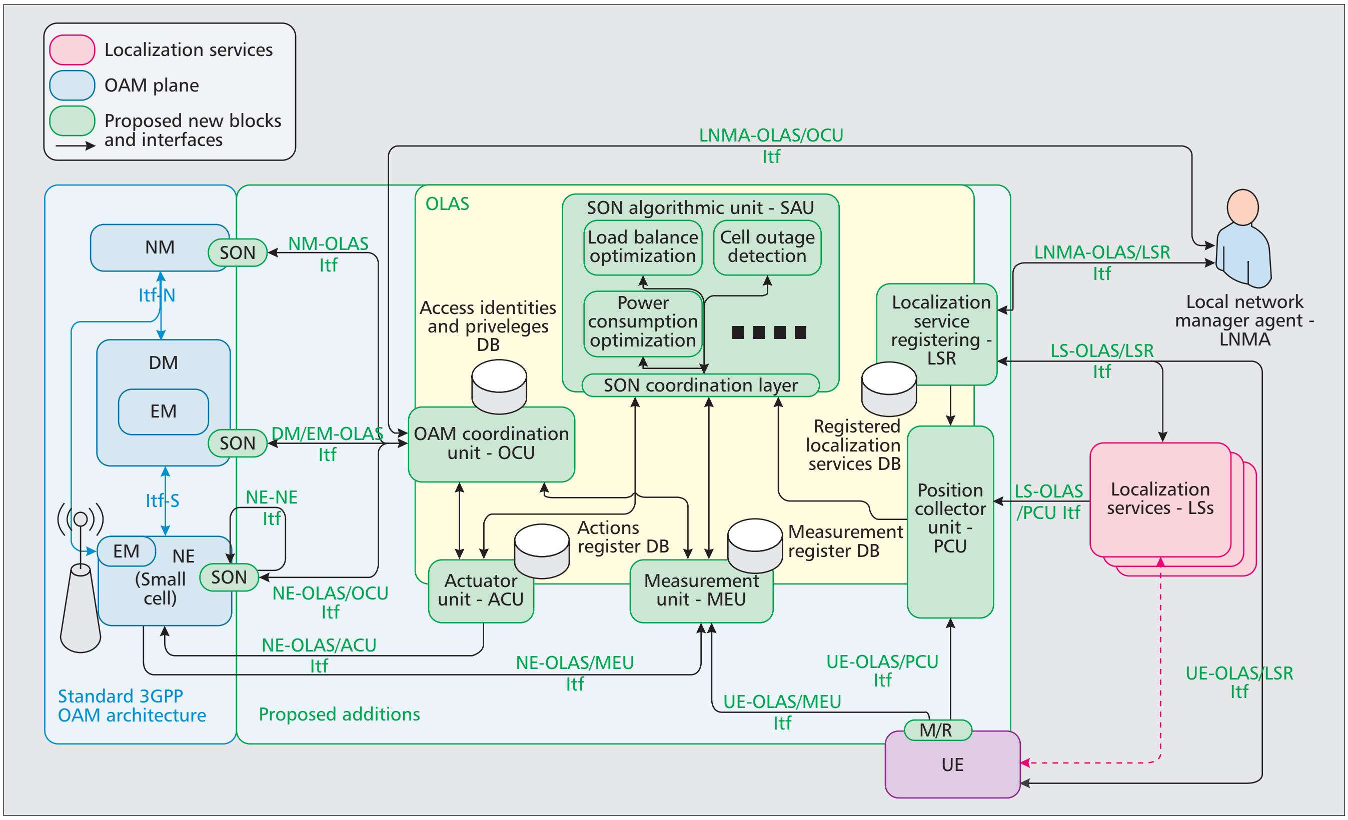


Figure 1. Functional OAM architecture.

adaptable to any particular indoor scenario and terminal distribution in order to take full advantage of the synergies between SON and positioning capabilities.

OAM System Architecture

Functional Architecture

The proposed OAM architecture is defined by several interrelated entities. Here, different approaches can be adopted in order to establish its functional scheme: *centralized* (where a unique entity is in charge of managing the rest of the elements); *distributed* (peer-to-peer); and *hybrid* (with a combination of the characteristics of the previous options; e.g., some mechanisms are completely local while others require coordination among distributed entities). The proposed solution follows a hybrid scheme. Even if this approach implies the need for a centralized OAM element, it is chosen as it allows easy reuse of classical centralized OAM architecture, while the implementation of distributed mechanisms is also supported.

Thus, the standard 3GPP OAM architecture [10] is maintained, adding new capabilities, functions, entities, and interfaces to it. The placement of SON functions in the standard OAM elements (network manager, NM; domain manager, DM; element manager, EM; and network element, NE) (Fig. 1) follows a scheme similar to that presented in [9]. Here the functions involving a specific subnetwork can be implemented at the DM layer. For functions involving more than one subnetwork (e.g., coordination with macrocell coverage), the function will reside on higher layers of the OAM hierarchy. Conversely, distributed SON functions would be placed at the EM layer and NE (e.g., small cells).

The level occupied by the tool/mechanism at the OAM architecture chain is directly related to the time span for monitoring/configuration and also the level of abstraction over the network elements [9] (Table 1).

However, even for the lowest layer standard centralized entity (DM), time spans (in the range of hours) are still large. Also, a DM usually operates non-overlapped subnetworks covering wide areas. Hence, a novel additional OAM functional block, the OAM location-aware system (OLAS), is proposed to support innovative location-based SON mechanisms. This new proposed centralized entity is to be implemented at the lowest levels of the OAM hierarchy, being in charge of managing the set of small cells of one specific indoor area.

The proposed architecture is shown in Fig. 1. Here, the standard OAM architecture is represented (blue left square) containing the described standard NM, EM/DM, and NE elements. These are connected by newly defined interfaces to the OLAS, which implements the following roles:

- It registers available localization services (LSs) and obtains location information from them.
- It implements location-aware SON functions.
- It acts as the coordinator for interaction between the OAM elements of the mobile network, location-aware SON algorithms, and LSs.

- It propagates the results of the location-aware SON algorithms to the OAM standard elements for their authorization to apply the decided commands in the network. Then these commands may be applied through standard OAM elements or directly to the devices by the OLAS itself depending on operator policies.

Also, additional *monitoring* and *reporting* functions (M/R) can be incorporated into the UEs, so they can directly report to the OLAS information on the network status or their location. This M/R capability can be part of the location-based applications present in the terminals (e.g., navigation app) or implemented by means of directly invoking functionalities in the terminal application programming interface (API).

The described OLAS roles are distributed in different functional elements, which allow better insight into the defined functionality.

SON algorithmic unit (SAU): Implements the local SON algorithms present in the system. It can contain multiple interdependent SON functions for self-configuration, self-optimization (e.g., load balancing or mobility robustness), and self-healing (e.g., cell outage detection and compensation). If multiple SON use cases are implemented, it would also be responsible for the proper coordination and trade-off between the different SON use cases and mechanisms by the SON coordination layer, its particularities being dependent on the specific use cases implemented. One benefit of the integration in the SAU of multiple SON mechanisms is that it supports the use of the same localization sources (as well as network indicators/measures) for the multiple use cases implemented, reducing the possibility of collisions generated by using different information sources, as well as allowing straightforward coordination between techniques.

Localization service registering (LSR): Is in charge of the incorporation and authentication of different sources of localization information into the registered localization services database. For an LS to be included, the main parameters for the information exchange with the OLAS have to be defined: IP address and format characteristics for communication with the LS. These parameters have to be compiled in a set of profiles to be used by the PCU in order to communicate with the different LSs available.

Position collector unit (PCU): Gathers the information coming from the LSs or the terminals registered in the system.

Actuator unit (ACU): Configures the network elements with the new parameters calculated by the location-aware SON algorithms, directly or by the standard OAM pile through the OAM coordination unit.

Measurement unit (MEU): Obtains information from the network elements by direct network element connection or through standard OAM elements using the OAM coordination unit. It is also in charge of the possible acquisition of direct network measurements from the UEs (received power levels, etc.).

OAM coordination unit (OCU): Serves as the interaction element between the OLAS and the OAM standard architecture. It translates the configuration orders coming from the ACU into commands for the operator's OAM tools and turns the OAM monitoring into a format usable by the MEU. Furthermore, it also supports the configuration of any of the OLAS functionalities by commands coming from the standard OAM architecture elements as well as by local network management agents.

	Task	Parameter abstraction	Time span
NM	Planning	Vendor-independent	Weeks/month
EM/DM	Network operation	Vendor-independent/ specific	Hours/days
NE	Element configuration	All parameters	Seconds/ minutes

Table 1. 3GPP standard characteristics of the OAM layers.

Local network management agent (LNMA): represents the specific personnel or administrator that may be required to manage the OLAS. The LNMA will have two main capabilities:

- It may register, via the LSR, new LSs to be used by the OLAS.
- It may alter the policies and/or functionalities of the OLAS via the OCU. This capability should be restricted through the permissions defined in the access identities and privileges database to avoid erroneous/malicious access.

Interface Protocols

According to the proposed architecture, the new main block OLAS introduces self-management at the local mobile network. Consequently, new interfaces, protocols, and applications should be implemented in order to coordinate this system with the rest of the OAM architecture as well as to measure and modify network devices.

NM-OLAS and DM/EM-OLAS: Used for the coordination between OLAS and the elements of the operators OAM core.

NE-OLAS: Exchanges monitoring and configuration messages between the small cells and the OLAS through three different interfaces:

- NE-OLAS/MEU focuses on monitoring and providing information about counters, alarms, KPIs, and so on to the MEU.
- NE-OLAS/ACU carries direct configuration commands or files to the NEs.
- NE-OLAS/OCU transports both monitoring and configuration messages when these cannot be directly sent/received to/from the NE by the OLAS blocks.

LS-OLAS: interfaces communication information from the LSs to the OLAS. These may be localization messages in order to support location-based SON functions (through the LS-OLAS/PCU Itf) or the procedure to register a new localization service in the LSR (through the LS-OLAS/LSR Itf).

LNMA-OLAS/OCU: allows (subject to operator permission) the configuration of the OLAS system by the LNMA. In turn, LNMA-OLAS/

Even if the considered scenarios are essentially local, the small cells are currently part of the mobile network operator infrastructure and make use of its radio spectrum. Therefore, as SON will alter the configuration of the cells, OLAS should remain on the operator's domain.

LSR serves for the manual registration of an LS by the LNMA.

UE-OLAS: logical connections send to the OLAS direct UE monitoring information (through the UE-OLAS/MEU interface) and UE provided localization information (by the UE-OLAS/PCU) that may be required for the SAU. For non-cellular external localization services, this interface would be UE-dependent; therefore, it may be only available for specific UE models such as smartphones.

All the interfaces connecting the OLAS with standard OAM architecture should follow the same standards as defined for 3GPP interfaces, being mainly based on TR-069 and XML [11]. LS-OLAS and UE-OLAS interfaces, however, are defined with elements that are independent of the mobile communications OAM network, as they are encapsulated on the mobile user/data plane, so any communication protocol (over IP) can freely be defined for these data flows.

3GPP LTE/LTE-A Femtocell Physical Implementation

The physical implementation of the proposed architecture for LTE/LTE-A systems is centered on the case where the deployed small cells are femtocells: HeNBs (home evolved NodeBs) [11] in 3GPP-LTE nomenclature. These are chosen because their limited capabilities, high vulnerabilities, and wide usage make this case especially challenging and comprehensive from an OAM perspective.

OAM-SON functionalities follow the standardized 3GPP architecture [10], with the novel addition of the OLAS, which implements local SON position-based functionalities. OLAS implementation can be *local* (if performed by hardware connected to the same LAN as the small cells) or *remote* (by external hardware connected to the system via the Internet). Remote solutions have high versatility in terms of using existing or leased equipment. However, the need for exchanging a high amount of information through the often limited network backhaul highly encourages the adoption of local implementations such as the one adopted here. Challenges for this approach include the need for additional OLAS on-site hardware, and its installation and maintenance, although the related cost is expected to be minimal over the total deployment expenses.

Figure 2 presents the local physical implementation, for LTE/LTE-A femtocell scenarios. The DM role is implemented by the HeNB management system (HeMS) [11]. The user and control planes of the HeNBs connect to the operator's core through the S1 interface, while the X2 interconnects the femtocells for distributed cooperation.

In this way, the defined logical links are implemented by physical connections as follows:

UE-OLAS and NE-OLAS: The information transmitted from the UEs to the OLAS is sent through the Uu interface (as part of its user/control plane) to the femtocell. This data (as well as the particular commands/information from the stations transmitted by the NE-OLAS interface) is then retransmitted through the LAN to the OLAS.

LS-OLAS: The interface used to transmit the UE location information (in case the positioning information is not directly obtained from the terminals) is implemented by the LSs through the Internet connection of the OLAS.

NM-OLAS and DM-OLAS: The coordination information between the OLAS and the operator's core (the elements of the standard highest OAM layers: DM and NM) is sent by the router through the backhaul to the operator's core.

The use of a LAN for exchanging data between the OLAS and the UEs greatly minimizes the traffic in the backhaul and the operator's core with reduced delay. This traffic local breakout has been envisaged by different manufacturers, standardization bodies [12], and other projects [7], and is a key factor for avoiding backhaul congestion by offloading signaling traffic, allowing reduced delays and time spans for the proposed location-aware functions.

Domain Responsibility and Security

This characteristic refers to the commercial/legal entity in charge of executing the different localization and/or SON functions. The responsible entities include the mobile network *operator*, the *user/administrator* of the local system, or a *third party*.

Even if the considered scenarios are essentially local, the small cells are currently part of the mobile network operator infrastructure and make use of its radio spectrum. Therefore, as SON will alter the configuration of the cells, OLAS should remain on the operator's domain.

Positioning information will be part of external systems out of the operator's domain, which may generate some difficulties. If the information coming from the localization system is accidentally or maliciously missing/erroneous, the impact on the mobile service provisioning generated by location-aware SON functions will be limited.

Moreover, the connections between the OLAS system and the external LSs shall avoid the disclosure (by direct access or simple traffic analysis) of network status information that may be sensitive. Thus, in the interaction between OLAS and the localization services, the extent, authenticity, accountability, and correctness of the information exchanged will be critical. Standardization on the LS-OLAS and LNMS-OLAS and related processes may be necessary to limit this issue.

Location-Aware SON Algorithms

In order to assess the usefulness of the proposed architecture, load balancing techniques are evaluated within the defined system. Here, the Power Traffic Sharing (PTS) algorithm presented in [13] has been selected as a consistent baseline method for our analysis. Within the proposed architecture, the algorithm (as well as any other SON functionality) is computed by the OLAS-SAU module. Based on a fuzzy logic controller (FLC), at the end of a time span t, this algorithm introduces variations on the transmitted power $\Delta TXP(i, t)$, for any small cell i in order to

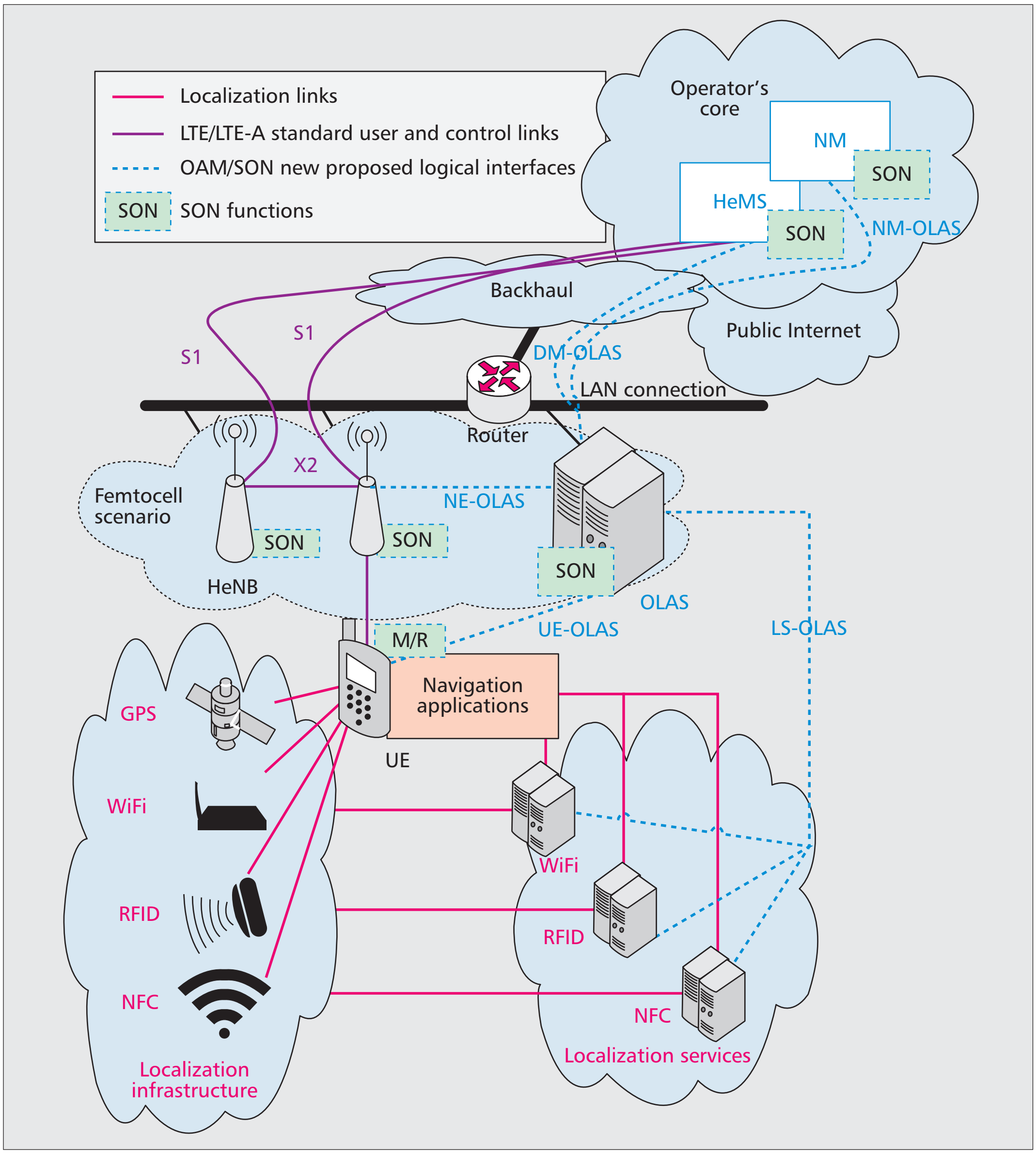


Figure 2. Proposed OAM architecture implementation for LTE/LTE-A femtocells.

In the interaction between OLAS and the localization services, the extent, authenticity, accountability, and correctness of the information exchanged will be critical. Standardization on the LS-OLAS and LNMS-OLAS and related processes may be necessary to limit this issue.

improve network performance via load balancing. In the proposed architecture, variations in the transmitted power are applied to the femtocells through the NE-OLAS/ACU interface, where the ACU is in charge of adapting the algorithm results to the required command format for the small cells.

In this load balancing system, traffic is distributed between cells by reducing/increasing their coverage area depending on the difference between each cell load ($load(i, t)$) and the average load of the remaining cells,

$$\overline{load(j,t)};\ \forall j \neq i.$$

Such load values are commonly indicators available from the base stations, and they are gathered through the NE-OLAS/MEU interface, following the specific commands for femtocell monitoring.

For this algorithm, $\Delta TXP(i, t)$ values tend to be conservative (low) as without information on users' positions, it is unknown whether each cell would gather more or fewer users after the power modification.

A novel location-aware SON algorithm (LPTS) based on PTS is proposed in Fig. 3, where the output of the FLC is weighted by a *convergence accelerator parameter*, $\alpha(i, t)$. The value of this parameter is calculated by the users-positions distribution analyzer depending on the UEs' spatial distribution on the selected cell and its neighboring cells. This block is implemented as part of the OLAS-SAU, and is the *user_positions*(t) input provided by the OLAS-PCU, which gathers the localization information directly from the terminals (through the UE-OLAS/PCU interface) or from localization services (LS-OLAS/PCU interface).

In this article, a simple $\alpha(i, t)$ function has been defined to assess the presented architecture: $\alpha(i, t)$ tends to double $\Delta TXP(i, t)$ values if the users' distribution changes and a concentration of users (more than 50 percent of them are

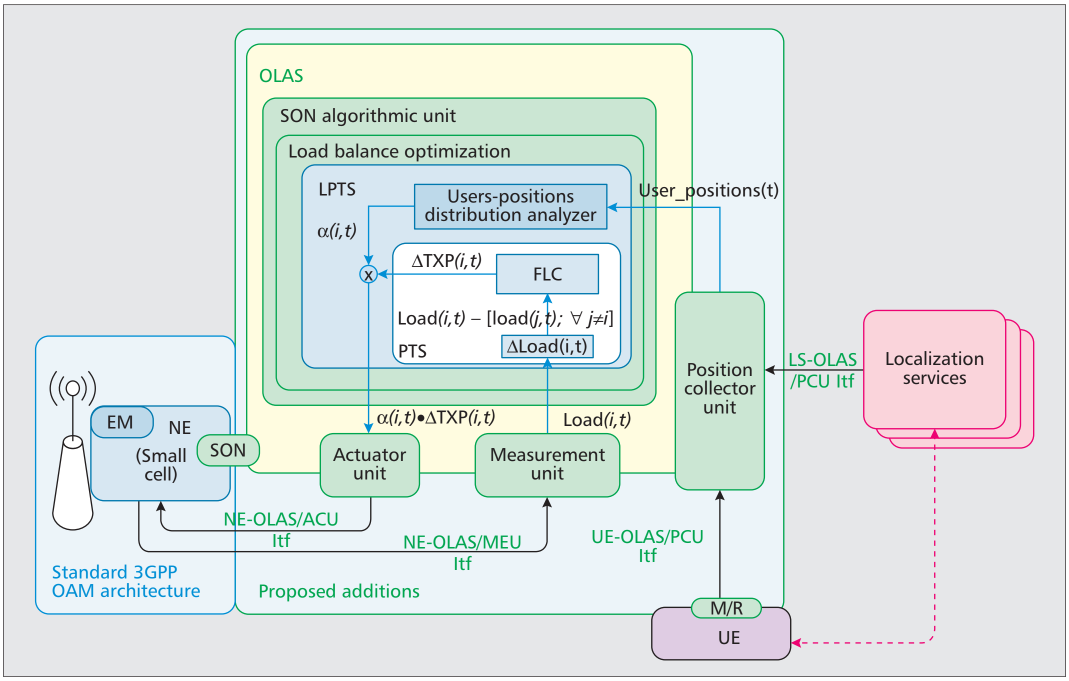


Figure 3. PTS and LPTS algorithms scheme.

> *The local character of the OLAS, from both functional and physical implementation perspectives, allows the reduction of the algorithm time span, that is, the time between two consecutive executions of the optimization algorithm and the period used to calculate the indicators of the network.*

located in less than 25 percent of the small cell area) is detected from one algorithm execution to the next. In consequence, high variations of power would be required in order to achieve load balance. In any other case, α(*i*, *t*) values are one; thus, LPTS works as does the classical PTS in this case.

Additionally, the local character of the OLAS, from both functional and physical implementation perspectives, allows the reduction of the algorithm *time span*, that is, the time between two consecutive executions of the optimization algorithm and the period used to calculate the indicators of the network.

Evaluation

The use of the localization information and the reduced time spans are both benefits provided by the architecture independent of the specific SON use case, providing an assessment of the capabilities introduced by the proposed system. Therefore, the presented location-aware load balancing algorithm, LPTS, is implemented following the proposed architecture. Results are compared with those of the previous baseline PTS algorithm in order to evaluate the usefulness of the proposed OAM architecture. Also, following the analysis presented in Table 1, the results for a "classic" value of 1 h are compared to those obtained for a time span of 15 min, which is supported by the proposed architecture.

Simulation Model

The evaluation of the presented use case algorithm is performed by the dynamic system-level LTE simulator presented in [14], using the same network parameters and user traffic. The simulated scenario consists of a 50 × 50 m office building with 5 floors, representing a typical corporate environment. Four small cells are deployed on the third floor where users' movements are based on the *random waypoint model* [15], where variable accumulation of users (in dense work areas, coffee talks, etc.) is also modeled.

As a relevant use case, a realistic daily users' position distribution is defined: during the first 4 h of the simulation, *small cell 1* is heavily loaded in its north area (i.e., a cell covering a working session room). During the next hours ($t > 4$ h), active users mainly gather in a narrow spot (at the bottom in Fig. 4) in the southwest area of neighboring *small cell 4* (i.e., the cell covering the coffee room).

Results

The simulations show the performance for the different self-optimization mechanisms and time spans presented:

- Without SON: without applying optimization.
- PTS-1hour: baseline PTS technique without localization information and using a long time span (1 hour). This is the case typically implemented in classic OAM architectures.
- PTS-15min: PTS using a reduced time span (15 min), supported by the proposed architecture.
- LPTS-1hour: location-based PTS mechanism where users' locations are provided by the proposed architecture. It uses a long time span.
- LPTS-15min: algorithm that makes use of both the reduced time span and the location awareness supported by the proposed architecture.

The main key performance indicator (KPI) analyzed is the unsatisfied user ratio (UUR). It is defined as a linear combination of the outage ratio (OR) and call blocking ratio (CBR), where

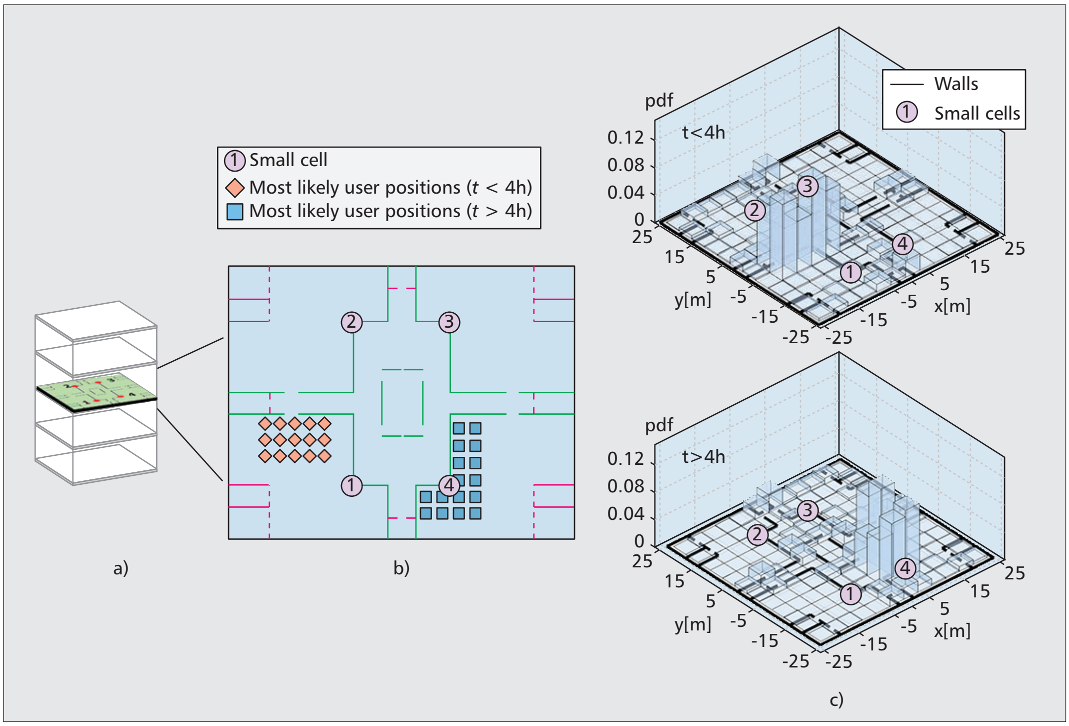


Figure 4. Spatial user distributions in the scenario (third floor): a) five-floor building; b) schematic of third floor active users' position distribution; c) third floor user densities.

OR is the probability that an existing network connection is in standby mode before it is finished (due to temporary lack of resources or bad signal-to-interference-plus-noise ratio, SINR), and CBR is the ratio of the number of blocked calls to the number of calls that attempt to access the network.

Figure 5 presents the simulation results, where the impact of the proposed architecture (availability of users' location and short periods to trigger the SON mechanisms) is observed in the evolution of the UUR indicator.

Before the change in the users' distribution (t <= 4 h), SON mechanisms exhibit better network performance (around half of the UUR value) than the situation without SON. Also, it is observed how reduced time span solutions provide better results for the same SON mechanism. On average, the 15 min solution enhances the overall network performance compared to the 1 h ones as it is able to achieve better adaptation to network temporal variations.

Once the variation in the user concentration has occurred (t > 4 h), the new condition leads to an even more degraded situation for the network without SON, with an increase in the average UUR from 12 to 26 percent. Meanwhile, the defined SON mechanisms are able to optimize and update configuration parameters in order to balance the network, reducing the UUR.

On one hand, the graph shows how *classic architectures* implementing SON (PTS-1hour) improves network performance compared to the case without SON. However, the adaptation to variations in users' distribution takes a long time (7 h). The reduced time spans (PTS-15min) supported by the proposed architecture accelerate that convergence period to 105 min. Furthermore, an improvement in UUR values is observed: a final UUR average of 11 percent for a 15 min span compared to 13 percent for time spans of 1 h. On the other hand, taking into account the use of localization, for equal time spans in SON mechanisms, LPTS converges to optimum values more than twice as fast as previous SON functions (PTS).

In this way, the *LPTS-15min* algorithm supported by the provided architecture converges around 9 times faster (taking 45 min) than classic SON mechanisms, also providing a reduction of around one third in UUR values. Additionally, it has to be noted that if user concentrations are continuously changing, it would be the only algorithm able to properly tackle these fluctuations, keeping proper network performance.

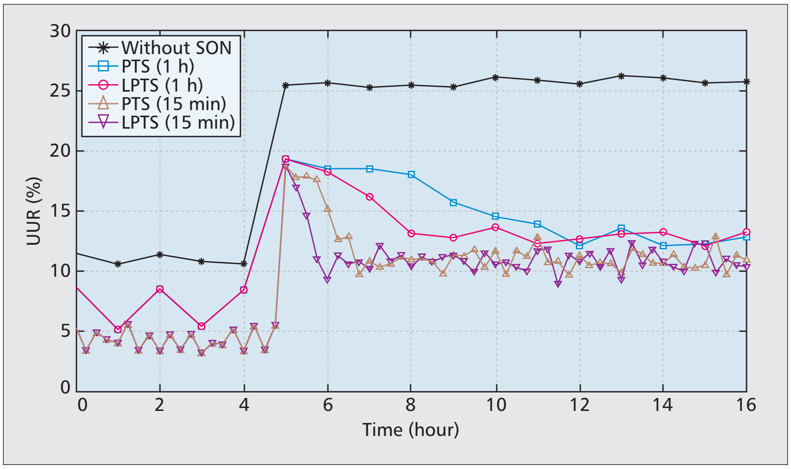


Figure 5. Evolution of UUR.

The impact of timing and location awareness in SON mechanisms has been analyzed in a key use case. The results show that the proposed architecture leads to a significant improvement in the response time by using location awareness and also advocates for the use of reduced time spans for these functions.

Conclusion

In this article, a novel OAM architecture extension for 3GPP-based self-organizing networks is proposed for medium/large indoor small cell scenarios with available positioning information. Its design is justified considering multiple architectural characteristics, features, and options. The achieved support for location-aware SON mechanisms is expected to lead to important improvements on mobile service performance at reduced cost.

Moreover, the impact of timing and location awareness in SON mechanisms has been analyzed in a key use case. The results show that the proposed architecture leads to a significant improvement in the response time by using location awareness and also advocates for the use of reduced time spans for these functions. This, together with the qualitative analysis of the OAM impact in terms of signaling overhead and delay, demonstrates that placing most of the SON functionalities at the lowest levels of the OAM hierarchy is the right decision.

Acknowledgments

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness within the National Plan for Scientific Research, Technological Development and Innovation 2008-2011, and the European Development Fund (ERDF) (IPT-430000-2011-1272). This work has also been partially supported by the Junta de Andalucía (Proyecto de Investigación de Excelencia P12-TIC-2905).

References

[1] "Boom in Mobile Data Creates Backhaul Urgency," white paper, Alcatel-Lucent, 2009.

[2] 3GPP TS 32.500, "Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements, v.11.1.0 (Release 11)," 2012.

[3] "Small Cells — What's the Big Idea; Femtocells Are Expanding Beyond the Home," white paper, Small Cell Forum Ltd., Feb. 2012.

[4] C. Brunner and D. Flore, "Generation of Pathloss and Interference Maps as SON Enabler in Deployed UMTS Networks," *Proc. IEEE VTC*, 2009, pp. 1–5.

[5] K. Kastell *et al.*, "Performance Advantage and Use of a Location Based Handover Algorithm," *Proc. IEEE VTC*, 2004.

[6] J. Steuer and K. Jobmann, "The Use of Mobile Positioning Supported Traffic Density Measurements to Assist Load Balancing Methods Based on Adaptive Cell Sizing," *Proc. 13th Int'l Symp. Personal, Indoor and Mobile Radio Communications*, 2002, vol.1, no. 1, 15–18 Sept. 2002, pp. 339–43.

[7] BeFEMTO, 2012; http://www.ict-befemto.eu/.

[8] NGMN requirement document, "NGMN Top OPE Recommendations," Sept. 2010.

[9] S. Hmlinen, H. Sanneck, and C. Sartori, *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*, Wiley, 2011.

[10] 3GPP TS 32.101, "Telecommunication Management; Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication Management; Principles and High Level Requirements," v. 10.0.0 (Release 10), 2012.

[11] 3GPP TS 32.593, "Telecommunication Management; Home Enhanced Node B (HeNB) Operations, Administration, Maintenance and Provisioning (OAM&P); Procedure Flows for Type 1 Interface HeNB to HeNB Management System (HeMS)," v. 10.2.0 (Release 10), 2011.

[12] 3GPP TR 23.829, "Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)," v. 10.0.1 (Release 10), 2011.

[13] J. M. Ruiz-Aviles *et al.*, "Fuzzy Logic Controllers for Traffic Sharing in Enterprise LTE Femtocells," *Proc. IEEE VTC-Spring '12*, Yokohama, Japan, May 2012.

[14] J. M. Ruiz-Aviles *et al.*, "Design of a Computationally Efficient Dynamic System-Level Simulator for Enterprise LTE Femtocell Scenarios," *J. Electrical and Computer Engineering*, vol. 2012, 2012.

[15] D. B. Johnson and D. A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," *Mobile Computing*, 1996, Kluwer Int'l. Series Eng. and Comp. Sci. 353, p. 153.

Biographies

Sergio Fortes (sfr@ic.uma.es) received his M.Sc. degree in telecommunication engineering from the University of Málaga, Spain, in 2008. He began his career being part of the main European space agencies (DLR, CNES, ESA) and Avanti Communications plc, where he participated in various research and consultant activities on broadband and aeronautical satellite communications. In 2012 he returned to the University of Málaga, where he is currently pursuing his Ph.D. focused on self-organizing networks for cellular communications.

Alejandro Aguilar-García (aag@ic.uma.es) graduated in telecommunications engineering in 2010 from the University of Málaga in the fields of telematics and communications. He started his career at Sony European Technology Centre in the Speech and Sound Group, participating in an existing video classification system based on audio and image features. He is currently working toward a Ph.D. developing novel SON mechanisms for small cells in mobile networks in the Communications Engineering Department of the University of Málaga.

Raquel Barco (rbm@ic.uma.es) holds an M.Sc. and a Ph.D. in telecommunication engineering from the University of Málaga. From 1997 to 2000, she worked at Telefonica in Madrid, Spain, and the European Space Agency in Darmstadt, Germany. From 2000 to 2003, she worked part-time for Nokia Networks. In 2000 she joined the University of Málaga, where she is currently an associate professor. Her research interests include satellite and mobile communications, mainly focusing on self-organizing networks.

Felix Barba Barba (felix.barba_barba@alcatel-lucent.com) is a technical Project Manager at Alcatel-Lucent in Madrid. He received a telecommunications engineer degree and an M.Sc. in telecommunication networks and telematic services from the Universidad Politécnica de Madrid, together with a degree in physics from UNED, Spain. He started his career in R&D in Telettra España (later Alcatel) on wireline access as a developer and project manager. His research interests include wireless techniques and other fields of innovation.

Jose Antonio Fernández-Luque (j.a.fernandez@alcatel-lucent.com) holds an M.Sc. in telecommunications from the University of Málaga. He started his career in Nokia R&D developing sensitivity analysis for automated troubleshooting. Later he joined TarTec, and later Optimi, where he worked on automated troubleshooting and optimization. He later moved to Nortel Networks (later Alcatel-Lucent), where he worked as an RF design and optimization engineer. He is currently an RF design leader in LTE deployments at Alcatel-Lucent in Madrid.

Alfonso Fernández-Durán (alfonso.fernandez@alcatel-lucent.com) is regional product line manager for 4G wireless at Alcatel-Lucent Spain. He received an M.Sc. and a Ph.D. in telecommunication engineering from Universidad Politécnica de Madrid, and an International Master's in management from EM Lyon. He began his career at Alcatel's Corporate Research Center, later becoming project manager for LMDS development in Alcatel's Wireless Access Division. He is currently involved in activities related to wireless indoor solutions and wireless small cell solutions.

ERRATA

In the September 2014 issue of *IEEE Communications Magazine*, vol. 52, no. 9, pp. 78–86, the article "Radio Propagation Path Loss Models for 5G Cellular Networks in the 28 GHz and 38 GHz Millimeter-Wave Bands" by Ahmed Iyanda Sulyman, AlMuthanna T. Nassar, Mathew K. Samimi, George R. MacCartney Jr., Theodore S. Rappaport, and Abdulhameed Alsanie had two errors.

The first error was in the legend of Fig. 1. The legend shows a path loss exponent "n_{best} = 4.5," but it should actually show "n_{best} = 3.7."

The second error was present in Table 1, row 7, starting with "TX Gain (dBi)." The last column shows a "1" when it should actually be "15."

Advertisers' Index

ADVERTISING SALES OFFICES

Closing date for space reservation: 15th of the month prior to date of issue

National Sales Office
James A. Vick
Sr. Director Advertising Business
IEEE Media
Email: jv.ieeemedia@ieee.org

Marion Delaney
Sales Director
IEEE Media
Email: md.ieeemedia@ieee.org

Susan E. Schneiderman
Business Development Manager
IEEE Tech Societies Media
Tel: (732) 562-3946
Fax: (732) 981-1855
Mobile: (732) 343-3102
Email: ss.ieeemedia@ieee.org

Northern California
George Roman
Tel: (702) 515-7247
Fax: (702) 515-7248
Cell: (702) 280-1158
Email: George@George.RomanMedia.com

Southern California
Patrick Jagendorf
Tel: (562) 795-9134
Fax: (562) 598-8242
Email: pjagen@verizon.net

Northeast
Merrie Lynch
Email: Merrie.Lynch@celassociates2.com
Tel: (617) 357-8190
Fax: (617) 357-8194

Jody Estabrook
Email: je.ieeemedia@ieee.org
Tel: (77) 283-4528
Fax: (774) 283-4527

Southeast
Scott Rickles
Tel: (770) 664-4567
Fax: (770) 740-1399
Email: srickles@aol.com

Midwest/Central Canada
Dave Jones
Tel: (708) 442-5633
Fax: (708) 442-7620
Email: dj.ieeemedia@ieee.org

Midwest/Ontario, Canada
Will Hamilton
Tel: (269) 381-2156
Fax: (269) 381-2556
Email: wh.ieeemedia@ieee.org

Texas
Ben Skidmore
Tel: (972) 587-9064
Fax: (972) 692-8138
Email: ben@partnerspr.com

Europe
Rachel DiSanto
Tel: +44 1932 564 999
Fax: +44 1 1932 564 998
Email: rachel.disanto@husonmedia.com

Germany
Christian Hoelscher
Tel: +49 (0) 89 95002778
Fax: +49 (0) 89 95002779
Email: Christian.Hoelscher@husonmedia.com

Currently Scheduled Topics

Topic	Issue Date	Manuscript Due Date
Internet of Things/M2M from Research to Standards	August 2015	January 15, 2015
Software Defined 5G Networks for Anything as a Service	September 2015	January 15, 2015
Social Networks Meet Next Generation Mobile Multimedia Internet	October 2015	May 15, 2015

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

Anritsu MD8475A Signalling Tester